Artificial Intelligence in Medicine (2004) **32**, 71–83

ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Data mining techniques for cancer detection using serum proteomic profiling

Lihua Li[a,*], Hong Tang[a], Zuobao Wu[a], Jianli Gong[a], Michael Gruidl[b], Jun Zou[b], Melvyn Tockman[b], Robert A. Clark[a]

[a]*Department of Radiology, College of Medicine, H. Lee Moffitt Cancer Center and Research Institute, University of South Florida, Tampa, FL 33612-4799, USA*
[b]*Department of Interdiciplinary Oncology, H. Lee Moffitt Cancer Center and Research Institute, University of South Florida, Tampa, FL 33612-4799, USA*

Received 29 August 2003; received in revised form 30 January 2004; accepted 9 March 2004

KEYWORDS
Proteomics; Cancer detection; Data mining; Statistical testing; Genetic algorithm; Support vector machine

Summary *Objective:* Pathological changes in an organ or tissue may be reflected in proteomic patterns in serum. It is possible that unique serum proteomic patterns could be used to discriminate cancer samples from non-cancer ones. Due to the complexity of proteomic profiling, a higher order analysis such as data mining is needed to uncover the differences in complex proteomic patterns. The objectives of this paper are (1) to briefly review the application of data mining techniques in proteomics for cancer detection/diagnosis; (2) to explore a novel analytic method with different feature selection methods; (3) to compare the results obtained on different datasets and that reported by Petricoin et al. in terms of detection performance and selected proteomic patterns. *Methods and material:* Three serum SELDI MS data sets were used in this research to identify serum proteomic patterns that distinguish the serum of ovarian cancer cases from non-cancer controls. A support vector machine-based method is applied in this study, in which statistical testing and genetic algorithm-based methods are used for feature selection respectively. Leave-one-out cross validation with receiver operating characteristic (ROC) curve is used for evaluation and comparison of cancer detection performance. *Results and conclusions:* The results showed that (1) data mining techniques can be successfully applied to ovarian cancer detection with a reasonably high performance; (2) the classification using features selected by the genetic algorithm consistently outperformed those selected by statistical testing in terms of accuracy and robustness; (3) the discriminatory features (proteomic patterns) can be very different from one selection method to another. In other words, the pattern selection and its classification efficiency are highly classifier dependent. Therefore, when using data mining techniques, the discrimination of cancer from normal does not depend solely upon the identity and origination of cancer-related proteins.

1. Introduction

Since the human genome project started about 10 years ago, a wealth of information about the sequences of individual genes has been revealed. There has been great progress in the construction of

*Corresponding author. Present address: Department of Radiology, College of Medicine, University of South Florida, 12901 Bruce B. Downs Blvd., MDC 17, Tampa, FL 33612-4799, USA. Tel.: +1 813 979 6718; fax: +1 813 979 6724.
E-mail address: lilh@moffitt.usf.edu (L. Li).

0933-3657/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2004.03.006

physical and genetic maps of the normal human genome and in the identification of genes associated with human diseases [1]. With the near completion of the genome project, the focus of research is now moving to the task of identifying the structure, function, and interactions of the proteins produced by individual genes and their roles in specific disease processes. This shift is driven by research indicating that (1) the level of mRNA expression frequently does not represent the amount of active protein in a cell; (2) the gene sequence does not describe post-translational modifications of proteins, which may be essential for protein function and activity; (3) the study of the genome does not describe dynamic cellular processes [1]. A key area in the post-genome era is proteomics, the global analysis of cellular proteins [2]. The proteome has been defined as the complete set of proteins encoded by the genome. Recently, the term has been broadened to include the set of proteins expressed both in space and time. Proteomics originally was defined as the analysis of the entire protein component of a cell or tissue, and now encompasses the study of expressed proteins, including identification and elucidation of the structure-function relationship under healthy and disease conditions, such as cancer [3]. The application of proteomics can be expected to have a major impact by providing an integrated view of individual disease processes at the protein level [2].

1.1. Proteomics for cancer research

Recent improvements in the technology of protein analysis, and in particular the development of advanced bioinformatic databases and analysis software, have allowed the development of proteomics. Proteomics uses a combination of sophisticated laboratory techniques including two-dimensional gel electrophoresis, image analysis, mass spectrometry (MS), amino acid sequencing, and bioinformatics to quantify and characterize proteins. In particular, proteomics provides the possibility of identifying disease-associated protein markers to assist in diagnosis or prognosis and to select potential targets for specific drug therapy [2].

There are currently two major approaches in applying proteomics to identify new biomarkers for cancer research. The first one, a bottom-up approach, is mostly from the perspective of molecular biology. Efforts are focused on identifying and characterizing a specific biomarker/protein at the molecular level, investigating the relationship between the structure/function of the biomarkers and their roles in cancer development. With such understanding, methods and medicines are sought to diagnose/treat/prevent cancer. There have been some studies reported along this theme [4–9]. Unfortunately, progress in assessment of the clinical utility of these biomarkers has been slow, in part due to a lack of emphasis on translational research studies to fully explore the biological and clinical implications of their potential as diagnostic or prognostic biomarkers. Assessment of individual biomarkers has often met with disappointing results, as shown in Table 1. Few studies have simultaneously evaluated more than one candidate biomarker to enhance the ''test's'' diagnostic/prognostic sensitivity and specificity. Such studies have led to the belief that no single marker is likely to prove sufficiently predictive, therefore emphasizing the need for the development of panels of multiple diagnostic/prognostic markers [10,11]. The latter is thought to be necessary to address the robust heterogeneity demonstrated by most human cancers.

The second approach, a top–down one, is from the perspective of bioinformatics. In this approach, proteomic spectra of certain biomarkers, related to certain diseases like cancers, are generated by MS. Matrix-assisted laser desorption and ionization (MALDI) and surface-enhanced laser desorption and ionization (SELDI) are the two most frequently used techniques for collecting proteomics mass spectra. MALDI spectra contain proteins and fragments of the proteins generated from laser ablation. SELDI MS is a refinement of MALDI. Its underlying principle is surface enhanced affinity capture through the use of protein chips

Table 1 Single biomarker for cancer detection

Cancer type	Biomarker	Disadvantage	References
Prostate cancer	Prostate specific antigen (PSA)	Only 25–30% specificity; PSA production influenced by many factors	[10]
Ovarian cancer	Cancer antigen 125 (CA125)	PPV of less than 10–20% coupled with ultrasound	[37]
Breast cancer	CA15.3	Lack the adequate sensitivity (23%) and specificity (69%)	[28]

Table 2 Proteomic research for cancer detection using mass spectrum

Cancer type	Feature	Learning algorithms	Sensitivity (%)	Specificity (%)	References
Prostate cancer	Peak	Decision tree	83	97	[11]
Prostate cancer	Peak	Boosted decision tree	100[a]	100[a]	[22]
Astroglial tumor	Peak	Neural network (Neuroshell 2)			[23]
Liver cancer	SAM (significance analysis of microarrays) identified points	Neural network (EasyNN)	92	90	[18]
Ovarian cancer	Genetic algorithm selected points	Self-organizing clustering analysis	100	95	[16]
Breast cancer	Peak	Unified maximum separability analysis	93	91	[28]
Prostate cancer	Peak	Logistic regression analysis	93	94	[30]
Prostate cancer	Peak	Manual analysis	100	100	[25]
Prostate cancer	Peak	Manual analysis	100	100	[26]
Breast cancer	Peak	Manual analysis	100	96	[26]
Colon Cancer	Peak	Manual analysis	100	86	[26]
Breast cancer	Peak	Manual analysis	75–84	91–100	[27]

[a] (AdaBoost) 97 (Boosted decision stump feature selection).

consisting of chemical or biological surfaces that bind proteins. Both of these methods can profile a population of proteins in a sample according to the molecular weight and net electrical charge (m/z) of the individual proteins [12,13]. Analysis of large numbers of proteins sampled from different populations (normal, patients, various stages of cancer, etc.), generate profiles of mass spectra. These profiles can contain thousands of data points, and may reflect the pathological state of organs and aid in the early detection of the disease. To uncover differences in complex mass spectral patterns of proteins, higher order analysis is required. Efforts have been made to link mass spectral analysis with a high-order analytical approaches, such as data mining, using samples of known diagnosis to define an optimal discriminatory proteomic pattern, then to use this pattern to predict the identity of masked samples. The goal is to extract a proteomic pattern that is both sensitive and specific to a disease with high reproducibility. An advantage with this top–down approach is that it is not necessary to purify, identify, and develop antibodies to individual proteins to proceed to clinical assay development. Even though it will eventually be important to know the identity of the proteins to understand their functional role and to assess their potential as novel therapeutic targets, the top–down proteomic approach by mass spectroscopy coupled with heuristic pattern recognition/data mining algorithms may become superior to immunoassays as clinical analyte sensors for early detection of cancer/disease.

1.2. Data mining techniques applied to proteomics for cancer research

Cancer detection based on the application of data mining techniques to proteomic data has received a lot of attention in recent years [3,10,14–20]. The proteomic data are predominantly mass spectra of patients' tissue cells, blood, serum, or other body fluids generated by mass spectrometry, although in principle other forms of data could also be analyzed in a similar manner. A mass spectrum contains information about proteins and their fragments [12,13,21]. The mass spectrum data present a curve with peaks and valleys, where the x-coordinate is the ratio of molecular weight to the net electrical charge for a specific organic molecule, with Dalton as unit, and the y-axis is the intensity (quantity) of signal for the same molecule.

Development and application of data mining algorithms to these proteomic data is an essential part in determining the clinical potential of a protein biomarker. Up to the present, several types of cancers have been studied with this approach, including ovarian, breast, prostate, liver, and colon cancer. Table 2 lists some of the research reported in recent years including the cancer type, features used, the learning algorithms and detection/diagnosis performance. The data mining techniques applied in these studies can be summarized as follows. Due to the fact that these studies were taken on different types of cancers with different data sets, it is inappropriate to make a direct comparison between these methods. Instead, it is a summary of research status.

1.2.1. Decision trees

Adam et al. [11] applied decision-tree learning to mass spectra of prostate cancer patients. They used Ciphergen SELDI(r) software for peak detection, and decision trees for classification using the intensity levels of the nine highest discriminatory peaks as features. This technique gave 96% accuracy, 83% sensitivity and 97% specificity. They also explored several bioinformatics models, including purely biostatistical algorithms, genetic cluster algorithms, support vector machines and decision classification trees, which gave accuracies between 83 and 90%.

Qu et al. [22] reported a boosted decision tree method for analyzing mass spectra to diagnose prostate cancer using the data of Adam et al. [11]. Their feature selection method was similar to that of [11]. Two new classifiers were developed, i.e. the AdaBoost classifier and the Boosted Decision Stump Feature Selection classifier. For the AdaBoost classifier, the sensitivity was 98.5% with a 95% confidence interval of 96.5—99.7%, and the specificity was 97.9% with a 95% confidence interval of 95.5—99.4%. For the Boosted Decision Stump Feature Selection classifier, a sensitivity of 91.1% with a 95% confidence interval of 86.9—94.6% and a specificity of 94.3% with a 95% confidence interval of 90.7—97.1% were reported.

1.2.2. Neural networks

Ball et al. [23] applied a three-layer perceptron artificial neural network (ANN) (Neuroshell 2) with a back propagation algorithm to analyze mass spectra for predicting astroglial tumor grade (1 or 2). A prototype approach was developed that uses a model system to identify mass spectral peaks whose relative intensity values correlate strongly to tumor grade. With a three-stage procedure, they screened a population of approximately 100,000—120,000 variables and identified two ions (m/z values of 13,454 and 13,474) whose relative intensity patterns were significantly reduced in high-grade astrocytoma. The accuracy achieved was between 83 and 100% for predicting tumor grade, however, the sample size for this study was only 12.

Poon et al. [18] used neural networks to discriminate hepatocellular carcinoma from chronic liver disease. Two hundred and fifty significant differentiating proteomic features were identified with significance analysis of microarrays (SAM). The ANN model was developed with EasyNN (Ver. 8.1; Stephen Wolstenholme). The development method was of the feed-forward type, and the networks were trained by weighted back-propagation. The ANN model was composed of three layers, one input layer, one hidden layer, and one output layer, with seven nodes in the hidden layer. They correctly classified 35 out of 38 hepatocellular carcinoma cases and 18 out of 20 chronic liver disease cases.

1.2.3. Clustering

Petricoin et al. [16] combined a genetic algorithm with self-organizing cluster analysis for identifying ovarian cancer. They reported an optimum discriminatory pattern for ovarian cancer, which was defined by the amplitudes at five key m/z values 534, 989, 2111, 2251 and 2465. A sensitivity of 100%, with 95% confidence interval of 93—100%, and a specificity of 95%, with 95% confidence interval of 87—99% were reported.

The same technique was also applied to the diagnosis of prostate cancer [17]. The amplitudes at seven key m/z values 2092, 2367, 2582, 3080, 4819, 5439, and 18,220 defined the optimum discriminatory pattern for prostate cancer. They correctly predicted 36 out of 38 patients with prostate cancer, resulting in a 95% sensitivity with 95% confidence interval of 82—99%; and 177 out of 228 patients were correctly classified as having benign conditions, that is, 78% specificity with 95% confidence interval of 72—83%.

Poon et al. [24] applied a two-way hierarchical clustering algorithm to differentiate hepatocellular carcinoma from chronic liver disease. Two hundred and fifty significant differentiating proteomic features identified with SAM were subjected to two-way hierarchical clustering analysis. However, they did not report sensitivity, specificity, or accuracy.

1.2.4. Manual analysis

Several investigators analyzed mass spectra data using the Ciphergen System software, combined with manual visual inspection. The Ciphergen System software was used to detect protein peaks, and then visually differentiate mass spectra of cancer patients from those of non-cancer people according to the protein peaks.

Hlavaty et al. [25] used the Ciphergen System software to detect peaks in the mass spectra and found that a 50.8 kDa protein peak was present in all 36 prostate cancer samples, but not in any of the twenty healthy people.

Watkins et al. [26] used the same method to detect breast, colon and prostate cancer. They correctly identified 41/41 (100%) breast cancer cases and ruled out 27/28 (96%) of the non-cancer cases. For colon cancer, they correctly identified 43/43 (100%) cancer cases and ruled out 24/28 (86%) non-cancer cases. For prostate cancer, their results were the same as that of [25].

Sauter et al. [27] analyzed the mass spectra data for nipple aspirate fluid over a 5—40 kDa range, from twenty breast cancer patients and thirteen healthy people. They identified five proteins. The most sensitive and specific proteins were 6500 and 15,940 Da, found in 75—84% of cancer samples but in only 0—9% healthy people.

1.2.5. Statistical method

Li et al. [28] used the ProPeak package, which provides an analysis module based on unified maximum separability analysis algorithm (UMSA). They achieved a sensitivity of 93% and a specificity of 91% for breast cancer detection with bootstrap cross-validation.

Valerio et al. [29] studied the mass spectra of thirteen pancreatic cancer patients, nine chronic pancreatitis patients and ten healthy people. Using statistical χ^2-test, they found unique protein peaks for each of the three groups; however, they did not report the sensitivity, specificity, or accuracy of their method.

Cazares et al. [30] applied mass spectrometry for prostate cancer diagnosis. They used Ciphergen Peaks 2.1 software for peak detection and a logistic regression analysis method for classification. A sensitivity of 93% and specificity of 94% were reported.

2. Materials and methods

This research takes the top—down approach by using serum proteomic profiling. Serum SELDI spectra data from patients and a healthy screening population were used as input. The output separates cancer cases from non-cancer screened controls.

2.1. Data

Three serum SELDI MS data sets were used in this research to identify serum proteomic patterns that distinguish the serum of ovarian cancer cases from non-cancer controls. The data sets were downloaded from a public website: http://clinicalproteomics.steem.com. As explained on the website, Dataset I (Ovarian, 16 February 2002) was collected using the H4 protein chip, and includes 216 total samples—100 controls, 100 ovarian cancer, and 16 benign, in which the spectra were exported with the baseline subtracted, and therefore negative intensities were observed in the data. Due to the discontinuation of the H4 chip, the WCX2 chip was chosen as a replacement in the generation of Dataset II (Ovarian, 03 April 2002), which has the same kind of samples as Dataset I. Again these samples were processed by hand and the baseline was subtracted creating the negative intensities. Dataset III (Ovarian, 07 August 2002) was also collected using WCX2 protein array, but a new set of ovarian samples was used. The sample set included 91 controls and 162 ovarian cancers. The entire process of applying the samples to the chips was done using a robotic instrument. The SELDI MS data for each case is an ASCII file containing 15,155 points of m/z values with corresponding intensities. Fig. 1 shows three examples of serum mass spectrum data, which are from a patient with biopsy-proven cancer, a benign case and a control case respectively. The x-axis is the ratio of the molecular weight to the net electrical charge, and the y-axis is the signal intensity. The distribution of samples for the three data sets is shown in Table 3. All samples in each data set are divided into cancer and non-cancer (including control and benign cases) classes in this study.

2.2. Methods

2.2.1. Feature extraction

As described in the previous section, the data size of the protein spectra obtained by SELDI is 15,155 points for each case. It is impractical to use all of these data as the input features to the classification because (a) some data points may contain noise and therefore may increase errors in classification; (b) a large number of features increase computational need; (c) it is difficult to define accurate decision-boundaries in a large dimensional space. The feature extraction process here is to select the most significant points of SELDI data as the features for cancer detection. By training, it tries to remove irrelevant and/or redundant data points (features) from the data (feature) set, and finds the minimal size subset of data points as features that carries enough information to perform an efficient pattern classification. However, it is a NP-hard problem [31]. In our case with a size of 15,155 data points, the size of the features subset is $2^{15,155}$. Exhaustively examining all the subsets is computationally prohibitive. Using a greedy search, such as hill-climbing would have the attendant the risk of being stuck in a local maximum.

Feature selection methods can be classified into two categories [31]. If the feature selection process does not involve a learning algorithm, it is a *filter* approach; otherwise, it is a *wrapper* approach. One of the main differences between these two methods is the evaluation method. In the filter approach, Euclidean distance measures, information measures, dependency measures, and consistency measures are usually used; in the wrapper approach, the

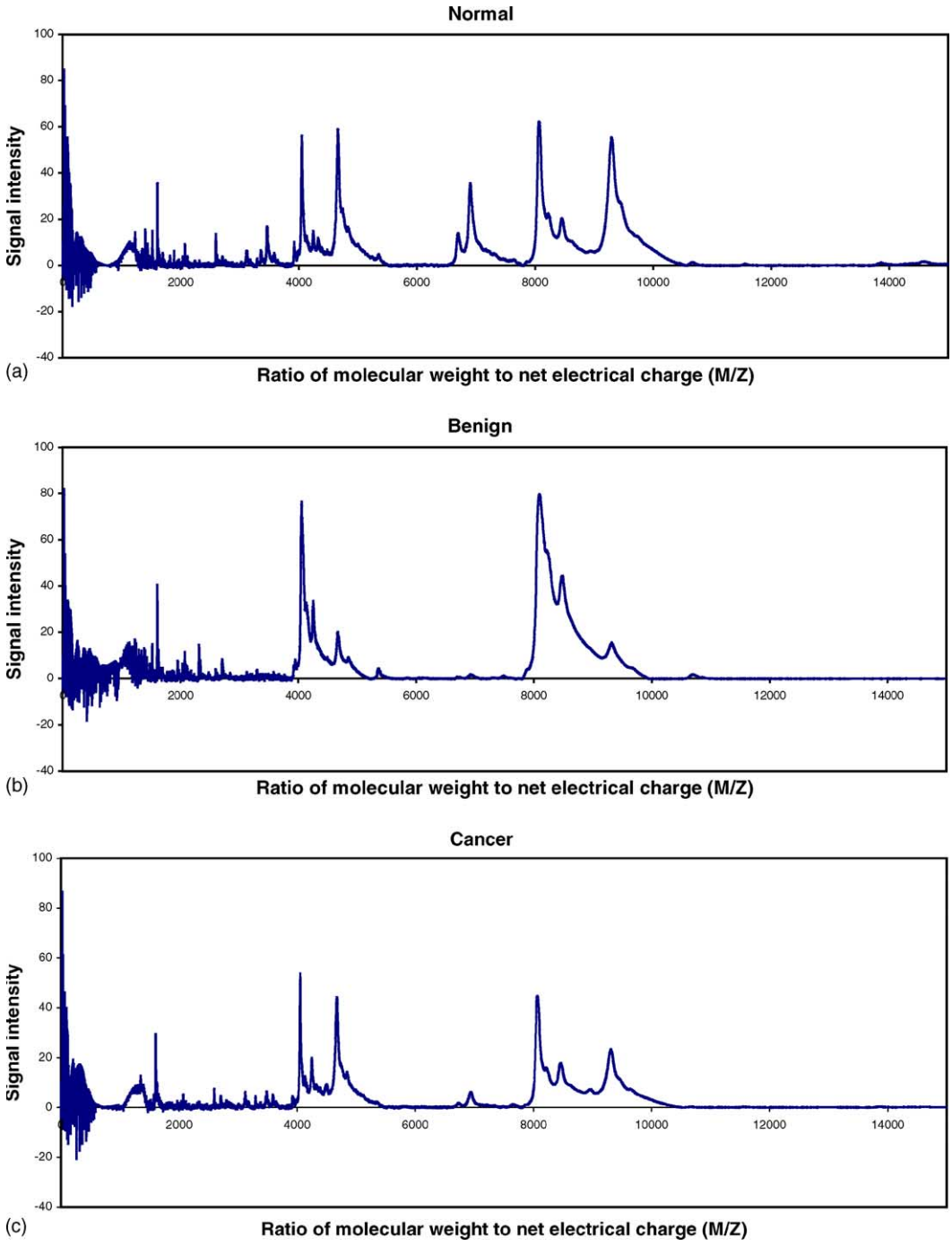

Figure 1 SELDI serum mass spectra from a control case, a benign case, and a patient with biopsy-proven cancer respectively.

Table 3 Distribution of samples

Datasets	Number of cancer cases	Number of control cases	Number of benign cases
Dataset I Ovarian, 16 February 2002	100	100	16
Dataset II Ovarian, 03 April 2002	100	100	16
Dataset III Ovarian, 07 August 2002	162	91	0

classifier error rate is used. The main advantage of the filter approach is its efficiency. Usually, it is much faster than the wrapper approach, but its accuracy is lower than the wrapper approach. Both of these feature selection approaches were explored in this study.

2.2.1.1. *Filter approach with statistical testing (ST)*

The filter approach to feature selection in this study was implemented by ST using a distance measure defined as follows

$$Y = \frac{|\mu_1 - \mu_2|}{\sqrt{\sigma_1 \times \sigma_1 + \sigma_2 \times \sigma_2}} \quad (1)$$

where μ_1 and μ_2 are the arithmetic means for the intensities at each m/z point of the cancer and non-cancer groups, respectively. σ_1 and σ_2 are the standard deviations of the corresponding intensities at each m/z point for both cancer and non-cancer groups.

The N m/z points with largest Y values are chosen as features, where N is the number of features. For this study, ten features are selected because our analysis shows that a feature size of ten is large enough for classification purpose [32].

2.2.1.2. *Wrapper approach with genetic algorithm (GA)*

Using a GA to select the optimal subset of features shows the robustness of the wrapper approach [33]. It provides the best individual for each generation. In other words, it is an anytime algorithm which allows us to allows a balance to be achieved between the quality of the selected features and computation time.

Initially, the genetic algorithm generates subsets (populations) of features uniformly distributed in the feature space. The learning algorithm evaluates each individual in the current population and assigns a fitness value to it. According to the fitness values, the genetic algorithm performs evolutional operations (selection, cross over and mutation) on the current population and produces a new generation whose individuals would have higher fitness values. In this way, the genetic algorithm explores the entire feature space in parallel, and the final result is globally optimal. In the experiments, we have set three stopping criteria: (a) the maximum number of generations is reached, (b) a fitness value of 1 (100%) is obtained, (c) the quality of the best individual meets the requirement. Fig. 2 shows the wrapper approach we used.

The GA algorithm used in feature selection is Genetic Algorithm Optimized for Portability and Parallelism (GALOPPS) R 3.2 by Erick D. Goodman, Michigan State University (http://garage.cps.msu.edu/software/software-index.html). It not only integrates most operators emerging in recent years, such as multi-field mutation operator, but also supports different coding methods. The fitness value is the classification accuracy determined by the support vector machine (SVM) classifier on individual leave-one-out dataset.

By studying the mass spectra, to reduce the computational load, we assume that the data with index above 10000 are considered irrelevant. GA chromosome is coded as a vector of integers whose range is [0—9999]. Because the m/z values on the x-axis of the spectra are real numbers, they need to be mapped into integers. Here the mapping simply uses their index. For example, the ith m/z value maps into (i—1), and so on. The parameters values set for GA are as follows:

- population size: 200;
- maximum number of generations: 8000;
- number of features: 10;

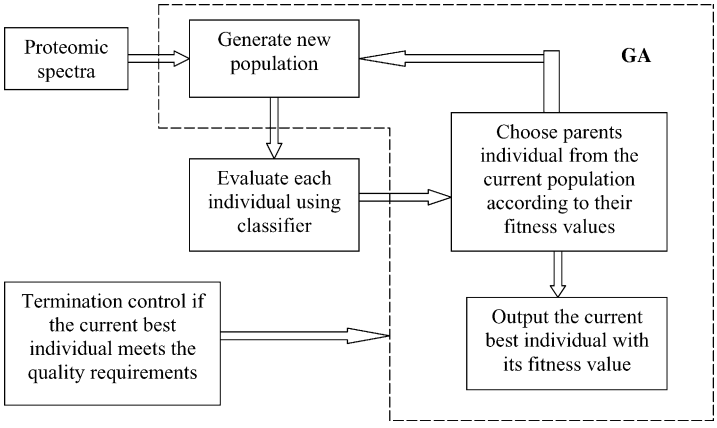

Figure 2 A wrapper approach to feature selection with GA.

- probability of crossover: 0.5;
- probability of mutation: 0.02.

The evolution operators used in the GA algorithm are stochastic universal sampling for selection, two-point crossover and multi-field mutation. For the purpose of comparison to the filter selection approach, the number of fields (features) was set to 10.

2.2.2. Classification method

A SVM was applied in this study as the classifier to discriminate the proteomic patterns identified by the two methods described above. SVM was first introduced by Vapnik [34] and has been recently proposed as a very effective method for regression, classification and general pattern recognition [35]. It is considered a good classifier because of its high generalization performance without the need to add a priori knowledge, even when the dimension of the input space is very high.

For a separable classification, given a training set of $S: x_1, x_2, \ldots, x_n$, where x_i represents the attribute vector of the examples and belongs to either one of two classes $y_i \in \{-1, 1\}$, a linear SVM finds the hyperplane leaving the points of the same class on the same side. The hyperplane can be represented as:

$$h(x) = \text{sign}(\langle w, x\rangle + b) = \begin{cases} 1, & \langle w, x\rangle + b > 0 \\ -1, & \langle w, x\rangle + b \le 0 \end{cases} \tag{2}$$

where w is a weight vector and b the threshold.

Each example is classified into class $+1$ or -1 based on which side of the hyperplane it lies on. The canonical representation of the separating hyperplane is obtained by rescaling the pair (w, b) into the pair (w', b'), such that

$$y_i(\langle w', x_i\rangle + b') \ge 1 \tag{3}$$

The distance of the closest point to the hyperplane equals to $1/w'$. These points are called support vectors.

The SVM approach is based on the structural risk minimization principle from statistical learning theory to find the optimal separating hyperplane (OSH) with the lowest probability of error. OSH minimizes the risk of misclassifying not only the examples in the training set but also the yet-to-be-seen examples of the test set. Vapnik shows that this goal can be equivalent to finding the hyperplane with maximum margin for separable training sets, i.e. SVM finds the hyperplane to separate the positive and negative training examples while maximizing the distance of either class from the hyperplane [34].

Since the distance of the closest point equals to $1/w'$, computing OSH is equivalent to solving the following quadratic optimization problem

$$\text{minimize} \quad \tfrac{1}{2}\langle w', w'\rangle \tag{4}$$

$$\text{Subject to} \quad y_i(\langle w', x_i\rangle + b') \ge 1, \quad i = 1, 2, \ldots, N \tag{5}$$

Using Lagrangian theory, we can transform the problem into the dual problem

$$\text{minimize} \quad \sum_i a_i - \frac{1}{2}\sum a_i a_j y_i y_j \langle x_i, x_j\rangle \tag{6}$$

$$\text{subject to} \quad y_i(\langle w', x_i\rangle + b') \ge 1, \quad i = 1, 2, \cdots, N \qquad a \ge 0 \tag{7}$$

where $a = (a_1, a_2, \ldots a_N)$ are the N non-negative Lagrange multipliers associated with the constraints (2).

For non-separable training set, the problem can be further transformed into the Soft Margin—Dual Lagrangian problem

$$\text{minimize} \quad \sum_i a_i - \frac{1}{2}\sum a_i a_j y_i y_j \langle x_i, x_j\rangle \tag{8}$$

$$\text{subject to} \quad \sum_i a_i y_i = 0, \quad i = 1, 2, \ldots, N \quad 0 \le a_i \le C \tag{9}$$

Here the pair (w', b') follows that

$$w' = \sum_i a_i y_i x_i \tag{10}$$

while b' can be determined from the Kuhn—Tucker Theorem condition as

$$a_i[y_i(\langle w', x_i\rangle + b) - 1 + \xi_i] = 0 \tag{11}$$

$$(C - a_i)\xi_i = 0 \qquad i = 1, 2, \ldots, N \tag{12}$$

where $\xi_i = 0 \quad (i = 1, 2, \ldots, N)$ for a separable training set.

Kernels were used for learning non-linear decision rules by mapping data into a richer feature space as $x \to \phi(x)$. Kernel is a function that calculates the inner product in some feature space as

$$K(x_1, x_2) = \langle \phi(x_1), \phi(x_2)\rangle. \tag{13}$$

Following are two common kernel functions used in SVM,

$$\text{polynomial kernel}: K(x_1, x_2) = (\langle x_1, x_2\rangle + b)^d \tag{14}$$

$$\text{Gaussian kernel}: K(x_1, x_2) = e^{-||x_1 - x_2||^2/2\sigma} \tag{15}$$

Because no big difference in performance was observed in classification using SVM with different

kernels, a linear kernel function (i.e. polynomial kernel with $d = 1$) was used in this study to save the training time and reduce the probability of overfitting. For numerical stability reasons, the kernel function was normalized by a factor of 20, which was selected empirically.

3. Results

We applied the proposed methods to the task of ovarian cancer detection using serum SELDI MS data. As listed in Table 3, three datasets were used for the training and testing; each of them contains biopsy proven ovarian cancer, control and benign samples. Because this is detection task, the serum samples in each dataset are divided into cancer and non-cancer groups in which the control and benign samples are grouped as a non-cancer set.

Two SVMs were trained using the features selected by statistic measure (SVM-ST) and GA algorithm (SVM-GA), respectively. Due to the limited size of each dataset, cross-validation within the original dataset was utilized to provide a nearly unbiased estimation of classification. In this study, leave-one-out cross validation is used for evaluation of cancer detection performance. For each classification, the true positive (TP), true negative (TN), false positive (FP) and false negative (FN) values are obtained. Accordingly accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) are calculated. Please note that the calculation of PPV and NPV follows Bayesian Theorem with a prevalence of 50 per 100,000, i.e. $P = 0.0005$ [36].

Table 4 lists the results of detection on three datasets with two different feature selection methods.

By varying the decision threshold of the SVM classifier, we can compute a receiver operating characteristic (ROC) curve, describing the trade-off between specificity and sensitivity. Using the area under the ROC curve (Az value), we can compare the performance of different classification tasks. In this study, the program ROCKIT provided by Charles E. Metz at the University of Chicago generated the ROC curves. Fig. 3 presents the ROC curves of the detections with Az values. Please note that the detection performance of SVM-GA on Dataset III is so good (100% accuracy) that its ROC curve could not be generated by the ROCKIT program (Az value is 1.0).

4. Discussions

The following observations resulted from this study: (1) overall, data mining techniques can be successfully applied to ovarian cancer detection with a reasonably high performance; (2) the classification using features selected by the genetic algorithm consistently outperformed that by filter approach feature selection. The GA based method is also less sensitive to the variation of datasets; (3) although the Dataset I and Dataset II include the same samples, the detection result on Dataset II is much better than that on Dataset I especially for detection with features selected by the filter approach. A reasonable explanation is that the SELDI data collected using WCX2 protein array is better than that by using H4 protein chip in terms of discrimination of proteomic profiling, even though both chips are from the same company (Ciphergen Biosystems, CA); (4) good detection results were obtained by both SVM-ST and SVM-GA methods on Datasets II and III, indicating that WCX2 proteomic array is a good profiling chip for discrimination.

Another important observation from our study is that for Dataset I (Ovarian, 16 February 2002), Petricoin et al. [16] reported a sensitivity of 100% (95% confidence interval 93—100), specificity of 95% (87—99), and PPV of 94% (84—99) by using a bioinformatic tool combining a GA with self-organizing cluster analysis for identifying ovarian cancer. Based on the same data set, we achieved a sensitivity of 79

Table 4 Detection results

Data set	TP	FP	FN	TN	Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)
Detection with features selected by filter approach with ST (SVM-ST)									
Dataset I Ovarian, 16 February 2002	79	23	21	93	79.0	80.1	79.6	0.20	99.99
Dataset II Ovarian, 03 April 2002	98	6	2	110	98.0	94.8	96.3	0.93	100.00
Dataset III Ovarian, 07 August 2002	160	3	2	88	98.8	96.7	98.0	1.48	100.00
Detection with features selected by wrapper approach with GA (SVM-GA)									
Dataset I Ovarian, 16 February 2002	96	6	4	110	96.0	94.8	95.4	0.92	100.00
Dataset II Ovarian, 03 March 2002	98	1	2	115	98.0	99.1	98.6	5.17	100.00
Dataset III Ovarian, 07 August 2002	162	0	0	91	100.0	100.0	100.0	100	100.00

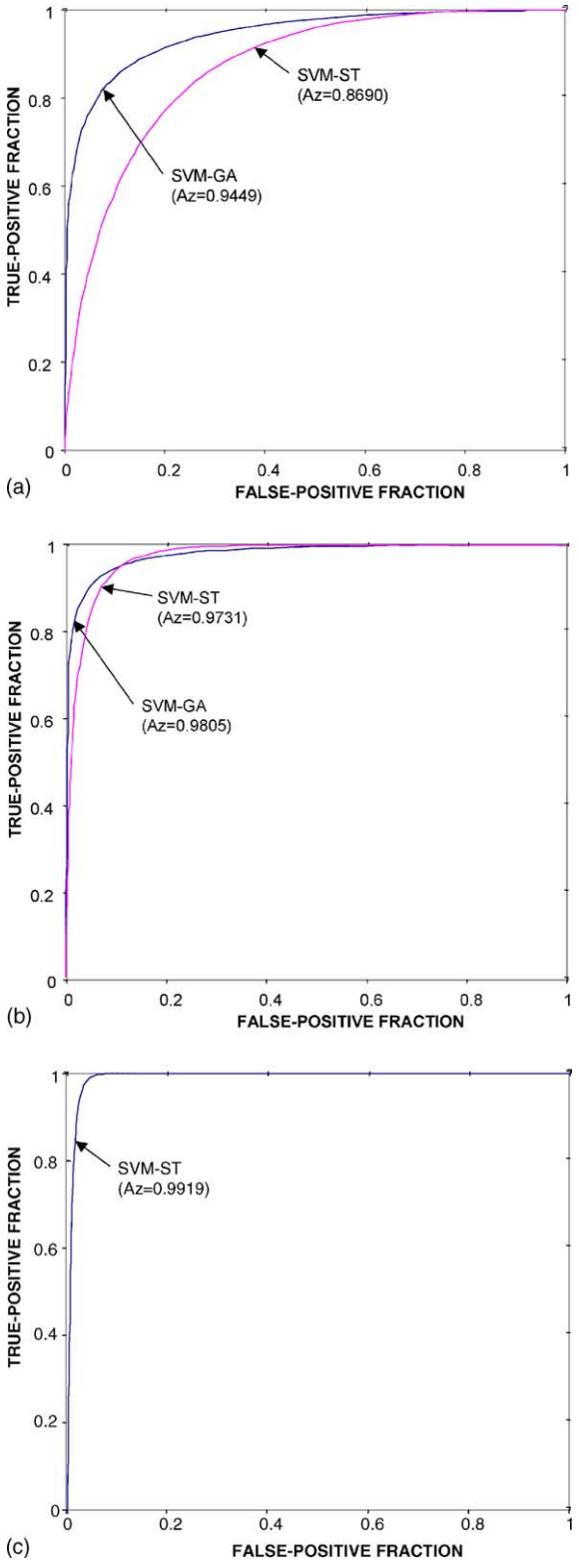

Figure 3 The ROC curves of SVM-ST and SVM-GA detections on: (a) Dataset I; (b) Dataset II; and (c) Dataset III (Az = 1 for SVM-GA on Dataset III).

and 96%, specificity of 80.1 and 94.8%, PPV of 0.20 and 0.92%, and NPV of 99.99 and 100% by using SVM as the classifier with two different feature selections, respectively. The performance difference results from the classification/feature selection methods as well as the evaluation methods. As described above, a leave-one-out cross validation was used here while in the Petricoin's study, the dataset was split into two subsets, one half (50 cancer and 50 non-cancer samples) was used as a training data while the second half (50 cancer and 66 non-cancer samples) was used as a test set.

The NPV is naturally high for screening applications. However, due to the low prevalence of ovarian cancer in the general population, the PPV is expected to be low. The PPV of 94% reported in [16] is not the PPV for real world screening. Instead it is the PPV for their experimental population composed of half cases and half controls. Using a more realistic prevalence of 50 per 100,000 for ovarian cancer for calculation, the PPV for their study would be about 0.99%. Both their study and ours suggest that the PPV for data set Ovarian, 16 February 2002 would be low if applied as a real world screening test. Again assuming a prevalence of 50 per 100,000, to obtain a PPV of 50%, one needs to reach a sensitivity of 100% and specificity of 99.95%; to obtain a PPV of 90%, one needs to get a sensitivity of 100% and specificity of 99.995%. Thus, to develop a screening tool for the general population, further study is needed to achieve better results from serum proteomic profiling of ovarian cancer.

It must be pointed out that at the proteomic level, there may be two types of 'biomarkers' that can be related to cancer. It could be that cancer results in the presence of specific protein(s), which is/are not present in the non-cancer environment. The cancer can be diagnosed by detecting the physical presence of this/these specific protein(s). Alternatively, the cancer may not lead to expression of a novel protein. Instead, it may change the complex proteomic pattern of the tumor-host microenvironment. In this case, the ''biomarker'' may be those normal host proteins that are aberrantly increased or decreased in abundance. This is an application where the data-mining techniques can be most helpful. A pattern analysis approach takes into consideration this gain or loss of global protein expression, not limited to any single protein molecule. Although the extracted pattern may not be able to lead to the physical protein identity for its every feature, it can still be used for cancer classification, as long as it achieves high accuracy. In this context, there may be more than one pattern that can discriminate a certain

Table 5 A comparison of detection results of SVM with different features

Methods	TP	FP	FN	TN	Sensitivity (%)	Specificity (%)	Accuracy (%)
SVM-ST	79	23	21	93	79.0	80.1	79.6
SVM-GA	96	6	4	110	96.0	94.8	95.4
SVM-REF	47	19	53	97	47.0	83.6	66.7

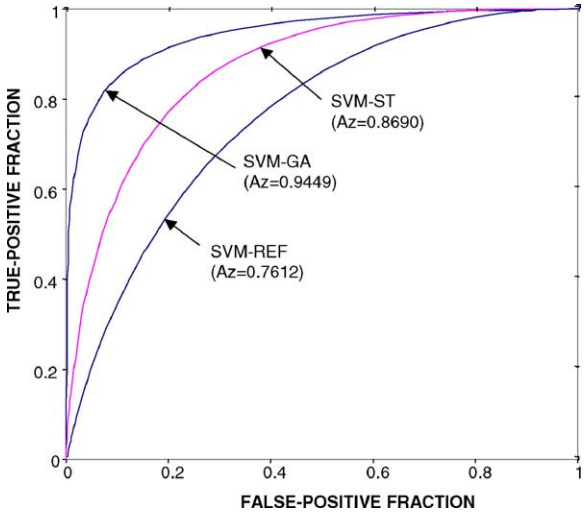

Figure 4 A comparison of SVM-ST, SVM-GA and SVM with ''reference'' features.

cancer. There may not be a single universal proteomic pattern for a specific cancer. It is based on this presumption that, even though the protein patterns are used as a detection/diagnosis paradigm, the data mining approach to cancer detection proceeds independently from the pursuit of the physiologic source and identity of these proteins. Therefore, in the context of data mining for cancer detection, discrimination of cancer from normal does not depend solely upon the identity and origination of cancer-related proteins. In fact, the identified discriminatory features (proteomic patterns) can be very different from one selection method to another. Also, the pattern selection and its classification efficiency are highly classifier dependent. To illustrate this fact, a cross evaluation was taken by using different features. Listed in Table 5 are the detection results of SVM with different features on Dataset I, in which SVM-REF is SVM trained and tested with the ''reference'' features identified in [16]. The ROC curves of detection on Dataset I using SVM-ST, SVM-GA and SVM with ''reference'' features are shown in Fig. 4. There is obviously a big drop in detection performance when the diagnostic patterns identified in [16] were ''transplanted'' to the SVM method. Accuracy is also worse than that obtained by the methods developed in this study.

5. Conclusions

Recent improvements in technology to detect, identify, and characterize proteins, particularly two-dimensional electrophoresis and mass spectrometry, coupled with development of bioinformatic databases and analysis software, make proteomics a powerful approach to identify new tumor markers. Nevertheless, large-scale studies will be necessary to validate these initial results and to determine clinical utility, assay reproducibility, and accuracy for diagnosis/prognosis of cancer.

This paper reviews the research of data mining techniques applied to proteomics for cancer detection/diagnosis. An SVM-based approach was applied to ovarian cancer detection using serum proteomic profiling MS data, in which statistical testing and genetic algorithm based methods were used for feature selection respectively. This study suggests that it is feasible to combine serum protein profiling with artificial intelligence learning algorithms to classify cancer samples from benign and/or normal controls with a top—down approach. Because these m/z values were found to be reproducibly detectable, only m/z values (a total of ten values in our study) are required to make an accurate detection. Their identities at the protein or molecular level are not necessary for classification purpose. However, due to the fact that knowing the identities of these discriminating substances is critical in understanding their biological role these peptide/proteins may have in the oncogenesis of ovarian cancer, and in identifying potential therapeutic targets, further research with bottom-up approach to purify, identify and characterize these protein/peptide biomarkers is important.

Further research will focus on feature design and selection. In this study all intensity values at the full m/z range are taken as potential features. Due to the high dimensionality, this method leads to a time-consuming and possibly suboptimal feature selection. The peaks of MS data have been used as the classification features in some studies, however the description of peaks was limited to their intensity values. More advanced features derived from each peak including its width, area, and height/width ratio may improve detection/diagnosis performance. In addition, validation on

extended database is also needed by including more cancer and control samples. How to integrate proteomic data and clinical data for improving early cancer detection is another very interesting and challenging research topic to be studied in future research.

Acknowledgements

This work is supported in part by a grant from NCI EDRN (U01 CA84973).

References

[1] Celis JE, Kruhoffer M, Gromova I, Frederiksen C, Ostergaard M, Thykjaer T et al. Gene expression profiling: monitoring transcription and translation products using DNA microarrays and proteomics. FEBS Lett 2000;480:2—16.
[2] Chambers G, Lawrie L, Cash P, Murray GI. Proteomics: a new approach to the study of disease. J Pathol 2000;192: 280—8.
[3] Srinivas PR, Srivastava S, Hanash S, Wright Jr GL. Proteomics in early detection of cancer. Clin Chem 2001; 47(10):1901—11.
[4] Alaiya AA, Franzen B, Auer G, Linder S. Cancer proteomics: from identification of novel markers to creation of artificial learning models for tumor classification. Electrophoresis 2000;21:1210—7.
[5] Alaiya AA, Franzen B, Hagman A, Dysvuk B, Roblick JU, Becker S et al. Molecular classification of borderline ovarian tumors using hierarchical cluster analysis of protein expression profiles. Int J Cancer 2002;98:895—9.
[6] Bergman A-C, Benjamin T, Alaiya A, Waltham M, Sakaguchi K, Franzen B et al. Identification of gel-separated tumor marker proteins by mass spectrometry. Electrophoresis 2000;21:679—86.
[7] Jones MB, Krutzsch H, Shu H, Zhao Y, Liotta LA, Kohn EC et al. Proteomic analysis and identification of new biomarkers and therapeutic targets for invasive ovarian cancer. Proteomics 2002;2:76—84.
[8] McDonald WH, Yates III JR. Shotgun proteomics and biomarker discovery. Disease markers 2002;18:99—105.
[9] Schmid H-R, Schmitter D, Blum P, Miller M, Vonderschmitt D. Lung tumor cells: a multivariate approach to cell classification using two-dimensional protein pattern. Electrophoresis 1995;16:1961—8.
[10] Adam B-L, Vlahou A, Semmes OJ, Wright Jr GL. Proteomic approaches to biomarker discovery in prostate and bladder cancers. Proteomics 2001;1:1264—70.
[11] Adam B-L, Qu Y, Davis JW, Ward MD, Clements MA, Cazares LH et al. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. Cancer Res 2002;62:3609—14.
[12] Bakhtiar R, Nelson RW. Mass spectrometry of the proteome. Mol Pharmacol 2001;60(3):405—15.
[13] Yates III JR. Mass spectrometry from genomics to proteomics. Trends Genet 16(1):5—8.
[14] Banks RE, Dunn MJ, Hochstrasser DF, Sanchez J-C, Blackstock W, Pappin DJ. Proteomics: new perspectives, new biomedical opportunities. The Lancet 2000;356:1749—56.
[15] Paweletz CP, Liotta LA, Petricoin III EF. New technologies for biomarker analysis of prostate cancer progression: laser capture microdissection and tissue proteomics, Urology 57(Suppl 4A):160—163.
[16] Petricoin Jr EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM et al. Use of proteomic patterns in serum to identify ovarian cancer. The Lancet 2002;359:572—7.
[17] Petricoin EF, Ornstein DK, Paweletz CP, Ardekani A, Hackett PS, Hitt BA et al. Serum proteomic patterns for detection of prostate cancer. J Ntl Cancer Inst 2002;94(20): 1576—8.
[18] Poon TCW, Johnson PJ. Proteome analysis and its impact on the discovery of serological tumor markers. Clin Chim Acta 2001;313:231—9.
[19] Wulfkuhle JD, McLean KC, Paweletz CP, Sgroi DC, Trock BJ, Steeg PS et al. New approaches to proteomic analysis of breast cancer. Proteomics 2001;1(10):1205—15.
[20] Wulfkuhle JD, Liotta LA, Petricoin EF. Proteomic applications for the early detection of cancer. Nature Rev Cancer 2003;3:267—75.
[21] Bakhtiar R, Tse FLS. Biological mass spectrometry: a primer. Mutagenesis 2000;15(5):415—30.
[22] Qu Y, Adam B-L, Yasui Y, Ward M, Cazares LH, Schellhammer PF et al. Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. Clin Chem 2002;48(10):1835—43.
[23] Ball G, Mian S, Holding F, Allibone RO, Lowe J, Ali S et al. An integrated approach utilizing aritificial neural networks and seldi mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. Bioinformatics 2002;18(3):395—404.
[24] Poon TCW, Yip T, Chan ATC, Yip C, Yip V, Mok TSK et al. Comprehensive proteomic profiling identifies serum proteomic signatures for detection of hepatocellular carcinoma and its subtypes. Clin Chem 2003;49(5):752—60.
[25] Hlavaty JJ, Partin AW, Kusinitz F, Shue MJ, Stieg A, Bennett K et al. Mass spectroscopy as a discovery tool for identifying serum markers for prostate cancer. Clin Chem 2001;47(10):1924—6.
[26] Watkins B, Szaro R, Ball S, Knubovets T, Briggman J, Hlavaty JJ, et al. Detection of early stage cancer by serum protein analysis. Am Aboratory 2001;32—6.
[27] Sauter ER, Zhu W, Fan X-J, Wassell RP, Chervoneva I, Bois GCD. Proteomic analysis of nipple aspirate fluid to detect biologic markers of breast cancer. Br J Cancer 2002;86: 1440—3.
[28] Li J, Zhang Z, Rosenzweig J, Wang YY, Chan DW. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. Clin Chem 2002;48(8):1296—304.
[29] Valerio A, Basso D, Mazza S, Baldo G, Tiengo A, Pedrazzoli S et al. Serum protein profiles of patients with pancreatic cancer and chronic pancreatitis: searching for a diagnostic protein pattern. Rapid Commun Mass Spectrum 2001; 15(24):2420—5.
[30] Cazares LH, Adam B-L, Ward MD, Nasim S, Schellhammer PF, Semmes OJ et al. Normal, benign, preneoplastic, and malignant prostate cells have distinct protein expression profiles resolved by surface enhanced laser desorption/ ionization mass spectrometry. Clin Cancer Res 2002;8(8): 2541—52.
[31] Liu H, Motoda H. Feature selection for knowledge discovery and data mining. Boston: Kluwer Academic Publishers; 1998.
[32] Hong T. Diagnosis of ovarian cancer based on mass spectrum of blood samples, MS Thesis. Department of

Computer Science and Engineering, University of South Florida; July 2003.
[33] Yang J, Honavar V. Feature subset selection using a genetic algorithm. Invited chapter. In: Motoda H, Liu H, editors. Feature extraction, construction, and subset selection: a data mining perspective. New York: Kluwer; 1998.
[34] Vapnik VN. Statistical learning theory, New York: Wiley; 1998.
[35] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge, UK: Cambridge University Press; 2000.
[36] Skates SJ, Xu F-J, Yu Y-H, Sjovall K, Einhorn N, Chang Y et al. Toward an optimal algorithm for ovarian cancer screening with longitudinal tumor markers. Cancer 1995; 76:2004—10.
[37] Woolas RP, Xu FJ, Jacobs IJ, Yu YH, Daly L, Berchuck A et al. Evaluation of multiple serum markers in patients with stage I ovarian cancer. J Ntl Cancer Inst 1993;85(21):1748—51.