

Proteomic Cancer Classification with Mass Spectrometry Data

Jagath C. Rajapakse, Kai-Bo Duan and Wee Kiang Yeo

BioInformatics Research Centre, School of Computer Engineering Nanyang Technological University, Singapore

Contents

Abstract	281
1. Mass Spectrometers: Fundamentals	282
1.1 Ion Source	282
1.2 Mass Analyzer	284
1.3 Ion Detectors	284
1.4 Tandem Mass Spectrometry (MS)	284
2. Analysis of MS Data	284
2.1 Baseline Correction and Noise Reduction	285
2.2 Normalization	285
2.3 Spectra Alignment	285
2.4 Peak Finding and Deconvolution	285
2.5 Peak Selection	285
2.6 Peptide Identification	285
3. Support Vector Machine (SVM)-Recursive Feature Elimination (RFE) Feature Selection and Classification	286
3.1 SVM	286
3.2 SVM-RFE	287
3.3 Numerical Experiments	288
4. Discussions and Conclusion	289

Abstract

The ultimate goal of cancer proteomics is to adapt proteomic technologies for routine use in clinical laboratories for the purpose of diagnostic and prognostic classification of disease states, as well as in evaluating drug toxicity and efficacy. Analysis of tumor-specific proteomic profiles may also allow better understanding of tumor development and the identification of novel targets for cancer therapy. The biological variability among patient samples as well as the huge dynamic range of biomarker concentrations are currently the main challenges facing efforts to deduce diagnostic patterns that are unique to specific disease states. While several strategies exist to address this problem, we focus here on cancer classification using mass spectrometry (MS) for proteomic profiling and biomarker identification. Recent advances in MS technology are starting to enable high-throughput profiling of the protein content of complex samples. For cancer classification, the protein samples from cancer patients and noncancer patients or from different cancer stages are analyzed through MS instruments and the MS patterns are used to build a diagnostic classifier. To illustrate the importance of feature selection in cancer classification, we present a method based on support vector machine-recursive feature elimination (SVM-RFE), demonstrated on two cancer datasets from ovarian and lung cancer.

The proteome is a highly dynamic entity whose variation reflects changes in physiological states in response to various stimuli. As such, the comparison of the appropriate proteomes has been increasingly useful in identifying diagnostic, prognostic, and therapeutic markers of disease and infection,^[1,2] as well as in evaluating drug toxicity and efficacy,^[3] leading to more effective and better tailored treatments.^[4,5] Proteomics also enables the characterization of various subtypes of disease, uncovering a number of novel potential drug targets.^[6] Cancer proteomics^[7,8] is an important subset of proteomics, involving the identification and quantitative analysis of differentially expressed proteins relative to healthy tissue counterparts at different stages of disease, from pre-neoplasia to neoplasia.^[9] Tumor-specific proteomic profiles are generated to better understand tumor development and progression while novel targets for therapy and novel markers for early diagnosis can be identified.^[10]

Generally, the earlier cancer is detected, the more favorable the prognosis for the patient.^[11,12] Unfortunately, in many cases, cancer is not diagnosed until cancer cells have metastasized and curative treatment is limited. This is especially applicable for cancers that present vague or no symptoms, or those that are relatively inaccessible to physical examination including breast, ovarian, liver, and lung cancer.^[5,11,13,14] Biomarkers that are specific and sensitive for a particular cancer type and detectable in high-risk patients or patients with early-stage cancer will be invaluable for detecting, staging, monitoring, and controlling the disease. With respect to cancer, protein biomarkers refer to substances or proteomic patterns that highlight the presence of cancer in the body; they can be compounds secreted by the tumor itself or as a result of a specific response of the body to the presence of cancer. Generally, biomarkers are measurable in tissues, cells or fluids, but to minimize the trauma and cost of screening, while maximizing utility, biomarkers that are measurable in serum,^[15] urine,^[16] or saliva^[17,18] are preferred.

Although the current state of cancer proteomics application is still largely limited to research, the ultimate goal is to adapt these applications for routine use in clinical laboratories. Apart from being used for the early detection of asymptomatic patients and diagnosis of symptomatic patients, biomarkers may potentially be used for the surveillance of individuals who have an increased chance of developing cancer. Also, biomarkers may be used to monitor patients with a previous medical history of cancer for recurrence.^[19] The clinical chemist provides patient-derived laboratory data to the physician who then incorporates information from various other assessments to make a diagnosis of the patient's condition. However, the raw laboratory data are of limited

value to the physician, and the clinical laboratory has to process the raw data to facilitate its interpretation. Clearly, the underlying presumption is that the methods used by the clinical chemist are accurate, consistent, and validated. The biological variability among patient samples as well as the huge dynamic range of biomarker concentrations are currently the main challenges facing efforts to deduce diagnostic patterns that are unique to specific disease states.^[20]

Many strategies are available for the identification of cancer biomarkers, including:

1. differential display of proteins, whereby normal and tumor lysates are compared and up- or downregulated proteins are identified^[21]
2. comparative analysis of secreted proteins and membrane fractions of different tumor cell lines to yield potential biomarkers^[6]
3. generation of unique protein profiles pertaining to specific tumors by using mass spectrometry (MS)^[22-28] (see figure 1).

Our focus in this review is on the classification of cancer by using MS data.

1. Mass Spectrometers: Fundamentals

MS^[29] involves the measurement of the mass-to-charge (m/z) ratio of ions and is increasingly being applied to sift through complex protein mixtures to find biomarker patterns that can be used for diagnosis, prognosis, or monitoring of disease.^[20] Two different types of instruments are mostly used for the majority of proteomics work: the matrix-assisted laser desorption ionization (MALDI)-time of flight (TOF) instruments and the electro-spray ionization (ESI)-tandem MS instruments.

Mass spectrometers contain at least three major parts: an ion source, a mass analyzer, and an ion collection/detection system. A sample is introduced into the mass spectrometer by using a direct insertion probe, direct infusion, or chromatographic separation interfaced with the MS instrument, which converts the components of a sample mixture to ions and then analyzes them on the basis of their m/z ratio.

1.1 Ion Source

The analysis of substances by mass spectrometers requires the formation of either positive or negative gas phase ions by a device referred to as the 'ion source' which produces ions by protonation ($M+H^+ \rightarrow MH^+$), cationization ($M+Cat^+ \rightarrow M Cat^+$), electron ejection ($M \rightarrow M^{+\bullet}+e^-$), electron capture ($M+e^- \rightarrow M^-$), deprotonation ($MH \rightarrow M^-+H^+$), or the transfer of a charged molecule from the condensed to the gas phase. The ESI and

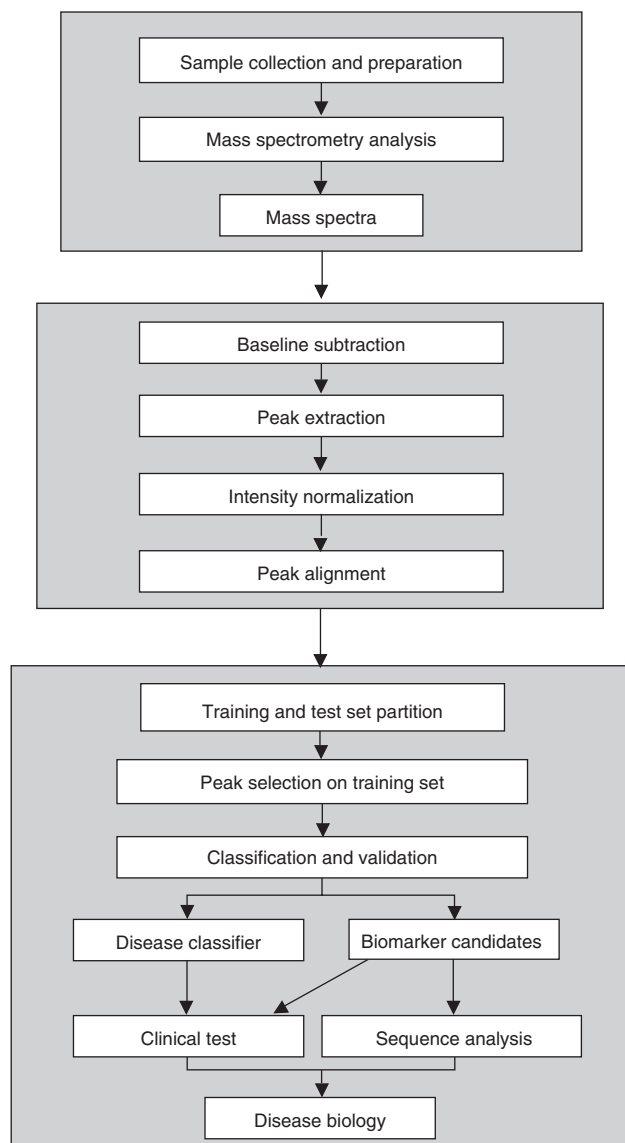


Fig. 1. An overview of basic steps involved in cancer classification and biomarker discovery with mass spectrometry.

MALDI methods make the ionization of large biomolecules possible and transform them to the gaseous phase. A high electric field is used in an ESI source to nebulize (spray out in a fine mist of droplets) the solution as it emerges from a needle; the small liquid droplets emerging from the needle contains both the volatile solvent and dissolved sample. While each droplet traverses a desolvation chamber, the solvent evaporates and causes its charge density to increase substantially. At a critical point, the charge density of the droplet exceeds the Rayleigh limit. This causes the ions to be ejected and then drawn into the mass analyzer. ESI commonly yields many peaks that correspond to the same species due to multiple charging.

In MALDI, samples are co-crystallized on the target plate after mixing with a matrix solution (an ultra-violet-absorbing compound). The co-crystal is irradiated by a pulsed laser beam, causing high-density energy to accumulate within it. The accumulated energy induces the samples and matrix molecules to vaporize. This results in proton transfer between the sample and matrix. This ionization process produces both positive and negative ions, depending on the nature of the sample. For peptides and proteins, the positive ions are formed by accepting protons as they are ejected from the matrix. Most of the peptide ions from MALDI are singly-charged because each peptide molecule tends to pick up a single proton.

Surface-enhanced laser desorption ionization (SELDI)-TOF is a technology involved in quantitative analysis of protein mixtures, which uses chemical or biological trapping surfaces that allow differential capture of proteins based on the intrinsic properties of the proteins themselves. Sample fluids are directly applied to the trapping surfaces, and any proteins in the sample fluids which have an affinity for the trapping surface will bind to it. The unbound proteins are removed with a series of washes. The bound proteins are laser desorbed and ionized for mass spectral analysis.

The role of SELDI-TOF in biomarker discovery primarily involves comparing the protein profiles from control and test samples and, thereafter, focusing on statistically significant differences.^[30] Although SELDI-TOF is able to rapidly deduce proteomic patterns from complex samples via differential comparison, this technology is still at an early stage of development. Therefore, SELDI-TOF also faces a number of limitations. One of the main limitations is the reproducibility of SELDI-TOF protein profiling experiments.^[31-34] Secondly, SELDI-TOF is limited in terms of sample resolution. Although it is particularly suited to proteins with a molecular mass below 30 kDa,^[35] SELDI-TOF poorly resolves higher molecular mass proteins. Thirdly, SELDI-TOF selects proteins based on chemical or biochemical properties of the chip array used. But the choice of chip array requires prior knowledge of the differences between the control and test samples so as to capture proteins with matching chemical or biochemical properties. Fourthly, there may be a bias of SELDI-TOF technology towards high-abundance proteins in the biological samples. This is particularly true for serum samples, which contain a huge variety of extremely high-abundance proteins resulting in the low-abundance proteins being out-competed for immobilization on the chip array.^[36] Lastly, the identification of proteins contributing to the SELDI-TOF proteomic patterns can be time-consuming^[35] since it is primarily designed to detect patterns rather than identify specific peaks in the spectra.

1.2 Mass Analyzer

Mass analyzers select ions over a particular m/z range. After the sample has been ionized at the ion source, the beam of ions is directed into a mass analyzer. The most common mass analyzers are the quadrupole ion trap, Fourier transform ion cyclotron resonance analyzer, and TOF mass analyzer.

Quadrupoles are four parallel rods with a direct current voltage and a radio frequency potential. The field on the quadrupoles determines which ions are allowed to reach the detector and, thus, quadrupoles function as a mass filter. As the field is imposed, ions moving into this field will oscillate depending on their m/z ratio and, depending on the radio frequency field, only ions of a particular m/z value can pass through the filter without hitting the rods. The m/z of an ion is therefore determined by correlating the field applied to the quadrupoles with the ion reaching the detector. A mass spectrum is obtained by scanning the radio frequency field.

The TOF mass analyzer measures the time it takes for the ions to fly from one end of the analyzer to the other and strike the detector. The speed at which the ion flies down the analyzer tube is inversely proportional to its m/z value. The smaller the m/z , the faster the ion flies. Apart from mass and charge, the arrival time of an ion at the detector is also dependent on the kinetic energy of the ion.

1.3 Ion Detectors

There are two ways in which the ion detector can generate a signal (current) from incident ion. The ion detector can either generate secondary electrons or induce a current generated by a moving charge. Among the detectors, the electron multiplier and scintillation counter are probably the most commonly used.

An electron multiplier detects ions with high sensitivity by extending the principle used with a Faraday cup. When an ion strikes the dynode surface of a Faraday cup, it causes several secondary electrons to be emitted. This change in charge creates a current in the cup and leads to a slight increase in the signal. An electron multiplier differs from a Faraday cup in that the former consists of a series of dynodes maintained at increasing potentials. The different dynodes attract secondary electrons which lead to a cascade of electrons.

The scintillation counter (photomultiplier conversion dynode) is similar to an electron multiplier; however, the scintillation counter consists of a phosphorus screen instead. When electrons strike the phosphorus screen, it releases photons. There is a photomultiplier which detects these photons. It operates with a

cascading action much like an electron multiplier. In addition, the photomultiplier tube is sealed in a vacuum, thus shielding it against contamination from the internal environment of the MS instrument.

1.4 Tandem Mass Spectrometry (MS)

Tandem MS (MS-MS) has been developed to induce further fragmentation of fragment ions. Such further fragmentation is accomplished by collisionally-generating fragments from a selected (parent) fragment ion by a process known as collision-induced dissociation (CID).^[37] This process selects an ion of interest with the mass analyzer and directs that ion into a collision cell. In the collision cell, the selected ion collides with an inert collision gas (typically argon). This results in the fragmentation of the ion. A daughter ion spectrum is obtained from the analysis of these new fragments. Tandem mass spectra generated as such is often abbreviated MSⁿ, where n refers to the number of generations of fragment ions being analyzed. For example, particular daughter ions (MS²) can be selected to undergo another round of fragmentation which results in granddaughter ions (MS³).

The selection of fragment ions to undergo MS-MS has been automated in some instrument control software. Such software is able to switch the mass analyzer between the usual full-scan and MS-MS modes to acquire tandem mass spectra. During the full-scan mode, the MS instrument is able to select the most intense ion and subject it to CID by automatically switching to MS-MS mode. Thereafter, the instrument automatically switches back to full-scan mode and proceeds to select the next most intense ion and subjects it to CID before switching back to full-scan mode. This approach is referred to as data-dependent MS-MS.^[38]

2. Analysis of MS Data

The output of the mass spectrometer is a mass spectrum, or chart, with a series of spike peaks each representing the ion(s) of a specific m/z value present in the sample. The heights of the peaks are related to the abundances of the ions in the sample. The height of the peaks and the m/z values provide a fingerprint of the sample. For protein samples, MS measures the m/z ratio of the ionized proteins (or protein fragments) and their abundances in the sample.

Recent advances in MS technology are starting to enable high-throughput profiling of the protein content of complex samples. For cancer classification, the protein samples from cancer patients and noncancer patients or from different cancer stages are analyzed through MS instruments and the MS patterns are used to build a diagnostic classifier. However, the raw mass spectra must

go through some basic preprocessing steps like baseline identification and subtraction, peak identification and extraction, intensity normalization, and peak selection, before they are used to build a cancer classifier. Figure 1 illustrates some basic steps involved in cancer classification and biomarker discovery with MS.

2.1 Baseline Correction and Noise Reduction

Every spectrum will have a base intensity level, known as baseline, which varies from spectrum to spectrum. When the spectrum (or a part of it) seems to be ‘lifted’ above the horizontal axis, baseline correction procedures must be implemented to bring the hanging baseline down. Wagner et al.^[39] used local linear regression to correct the baseline, and discrete wavelet transform (DWT) has been used in MS data for de-noising and data compression.^[40] Sauve and Speed^[41] used dynamic programming to improve calibration across multiple spectra and morphologic filters for baseline correction.

2.2 Normalization

It is necessary to normalize each m/z value in a spectrum so that it allows each value to be compared with others in a meaningful way. The most common method is to divide each data point in the spectrum by the total ion current (the summed intensities over all timepoints). Normalization is performed after the baseline correction.

2.3 Spectra Alignment

One of the most common problems related to data from MS is that there is often a ‘shift’ in the x-axis (m/z ratio) resulting in misalignment of profiles. The proper alignment of profiles is crucial for accurate data processing since multivariate analysis is sensitive to minor variations.^[42] While this problem has been mainly addressed with regards to chromatographic profiles,^[43-50] users of other spectral analytical techniques have also shown increasing interest.^[51-53] Various techniques have been used to address the problem of profile misalignment; of these the warping of spectra is one of the commonly used techniques. Warping compresses and stretches local segments of one spectrum so that the peaks are in horizontal alignment with a reference spectrum.^[46-49,52]

2.4 Peak Finding and Deconvolution

Due to the presence of spurious peaks, it is not desirable to rely on all the datapoints in the spectrum for establishing potential

biomarkers. Yasui et al.^[54] designed a peak finding algorithm which reduces the number of datapoints significantly. This was accomplished by first assessing, at each m/z point, whether or not the intensity of that point is the highest among its nearest $\pm N$ -point neighborhood set (with respect to the m/z axis); the optimal value of N was heuristically chosen.

2.5 Peak Selection

For MS data, the heights of thousands of peaks at different m/z values are the variables (features) that determine the dimensions of the instance space. The number of samples required for generalization increases exponentially with the number of variables. However, the number of variables far outnumbers the number of samples. When the number of dimensions reaches the hundreds, or even thousands, the computational time required for the classification algorithms can become prohibitive. Moreover, some of the variables are not discriminatory; in addition to the computational cost, irrelevant features may also cause a reduction in the accuracy of some classification algorithms.

Since it is often impossible to generate more sample data, the only viable option is to be very selective in the number of variables used for classification by identifying the most significant features and, at the same time, retain as much as possible of their class discriminatory information. Peak selection is exactly the feature/variable selection problem commonly addressed in machine learning.^[55,56] Some statistical and machine learning methods have been used for peak selection purposes, for example genetic algorithm,^[57] signal-to-noise ratio,^[58] and ROC curve criterion.^[59]

2.6 Peptide Identification

When selected peptide ions undergo further fragmentation in the collision cell, there are several bonds in peptides that could break to form new fragments. However, the most significant cleavages are along the peptide backbone. It is the amide bond between amino acid residues that most frequently breaks. To describe such peptide ion fragmentation during MS-MS, a widely accepted nomenclature is often used: when the bond between the carbonyl oxygen and the amide nitrogen is cleaved, the two ions thus formed are named ‘*y-ion*’ and ‘*b-ion*’. When doubly charged peptide ions are fragmented, both a *b-ion* and corresponding *y-ion* are formed since doubly charged ions are most likely to have charges at opposite ends of the molecule. A *y-ion* is a fragment in which the positive charge is retained on the intact C-terminus of the original peptide ion; a *b-ion* is a fragment in which the charge

is retained on the intact N-terminal portion of the original peptide ion.

There are two main ways to use MS-MS spectra for protein identification; first, one can try to interpret the spectra directly to infer a peptide amino acid sequence using the so-called ‘*de novo* sequencing process’ which derives the peptide sequences from given tandem mass spectral data of k ion peaks without searching against protein databases. Such methods will be useful if the sequence of interest contains molecular modifications that are not recorded in the databases or if it is a novel protein that has no matches in the databases. A number of *de novo* sequencing software programs are based on a graph theory approach^[60-63] in which a NC-spectrum graph is created (‘NC’ is an abbreviation for ‘N-terminal and C-terminal’, since the graph is built on the N-terminal *b*-ions and the C-terminal *y*-ions). A NC-spectrum graph consists of nodes and edges. Each peak in the MS-MS may be generated by several different types of ions. The nodes in the NC-spectrum graph attempts to represent the different types of ions for each peak in the MS-MS spectrum. Nodes are connected to one another with edges. Each edge actually represents the mass difference between two connected nodes. Therefore, the edge can often be labeled with the name of an amino acid if the mass difference matches the mass of that particular amino acid. Often, methods based on NC-spectrum graphs attempt to solve the peptide identification problem by finding the longest path in the NC-spectrum graph.

Recently, Frank and Pevzner proposed the PepNovo algorithm^[64] which is the most up-to-date improvement to the NC-spectrum graph approach. PepNovo is based on a scoring method which uses a probabilistic network and finds the highest-scoring path using a dynamic programming algorithm. PEAKS^[65] is one of the most popular *de novo* sequencing software packages, which finds a subset of best sequences of all possible combinations of amino acids for a specific precursor ion mass and then derives a consensus among the globally top-scoring sequences. Lutefisk^[66,67] uses the NC-spectrum graph approach and incorporates a FASTA-modified search algorithm (CIDentify) to queries sequence databases.

The second way of using MS-MS spectra for protein identification relies on a direct comparison of spectra with virtual theoretical spectra calculated from a sequence database. SEQUEST, an algorithm for identifying proteins by matching tandem MS data to database sequences, was introduced in 1994.^[68] Thereafter, several similar programs have been introduced to provide a relatively rapid assignment of MS-MS spectra to specific peptide sequences in databases. The MASCOT^[69] program uses the probability-

based MOWSE algorithm,^[70] precursor m/z information, and MS-MS fragment ion data to identify proteins from databases. MOWSE uses average properties of the proteins in the database to improve the sensitivity and selectivity of the identification. ProFound^[71] is an expert system for protein identification using Bayesian theory to rank the protein sequences in the database by their probability of occurrence.

3. Support Vector Machine (SVM)-Recursive Feature Elimination (RFE) Feature Selection and Classification

Support vector machine-recursive feature elimination (SVM-RFE) was originally proposed for gene selection,^[72] where a linear version of the popular SVM^[73,74] is used as the learning algorithm in a recursive procedure to select a subset of genes for cancer classification. In this section, we will present the usefulness of SVM-RFE for peak selection for cancer classification with MS data.^[75] For comparison, we also include the T-statistics feature selection method, which chooses a set of features that are most relevant to the concept under study. In this study the ‘goodness’ of the selected peak subsets is evaluated by the classification performance of a linear SVM classifier with only the selected peaks as input variables.

3.1 SVM

SVMs have been very popular for solving classification problems.^[73,74] SVMs construct an optimal hyperplane decision function in a so-called ‘feature space’ that is mapped from the original input space. The mapping ϕ is usually nonlinear and the feature space is usually a much higher dimensional space than the original input space. Let us use \mathbf{x}_i to denote the i th example vector in the original input space and \mathbf{z}_i to denote the corresponding vector in the feature space, $\mathbf{z}_i = \phi(\mathbf{x}_i)$.

Kernel is one of the core concepts in SVMs and plays a very important role. Kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$ computes the inner product of two vectors in the feature space and, thus, implicitly defines the mapping function (equation 1):

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) = \mathbf{z}_i \cdot \mathbf{z}_j \quad (\text{Eq. 1})$$

In what follows, we use a linear kernel, $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$.

For a typical classification problem with l training samples, $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_l, y_l)$, where y_i denotes the class label of x_i and $y_i \in \{1, -1\}$, finding the discriminant function $f(\mathbf{x}) = w \cdot \phi(\mathbf{x}) + b$ is

formulated by SVMs into the following optimization problem (equation 2):

$$\begin{aligned} \min_{\mathbf{w}, b, \xi_i} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^l \xi_i \\ \text{subject to} \quad & y_i(\mathbf{w} \cdot \mathbf{z}_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \end{aligned} \tag{Eq. 2}$$

where $C > 0$ is another predefined higher-level parameter, besides the kernel function parameters. This optimization problem is usually solved in its dual form^[74] (equation 3):

$$\begin{aligned} \min_{\alpha_i} \quad & \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j y_i y_j (\mathbf{z}_i \cdot \mathbf{z}_j) - \sum_{i=1}^l \alpha_i \\ \text{subject to} \quad & 0 \leq \alpha_i \leq C, \sum_{i=1}^l \alpha_i y_i = 0 \end{aligned} \tag{Eq. 3}$$

The weight vector \mathbf{w} and the hyperplane decision function can be expressed by using the dual variables α_i 's (equations 4 and 5):

$$\mathbf{w} = \sum_{i=1}^l \alpha_i y_i \mathbf{z}_i \tag{Eq. 4}$$

$$f(\mathbf{x}) = \sum_{i=1}^l \alpha_i y_i (\mathbf{z} \cdot \mathbf{z}_i) + b \tag{Eq. 5}$$

In the dual problem of SVMs, all the computation involving the input vectors is in the form of inner products of vectors in feature space. The discriminant function also can be expressed in inner products of feature space vectors. These inner products ($\mathbf{z}_i \cdot \mathbf{z}_j$) can be replaced by corresponding kernel computations $k(\mathbf{x}_i \cdot \mathbf{x}_j)$, which can be executed easily in the original input space. Thus, usually we do not need to know the explicit mapping function of ϕ , because it is implicitly defined by the kernel function that computes the inner product in the feature space. Similarly, we do not need to explicitly compute the weight vector \mathbf{w} . However, if a linear kernel is used, the decision function $f(\mathbf{x})$ is simply a linear function of \mathbf{x} and the weight vector of the linear function also can be explicitly computed as (equation 6):

$$\mathbf{w} = \sum_{i=1}^l \alpha_i y_i \mathbf{x}_i \tag{Eq. 6}$$

SVMs with linear kernels are often referred to as linear SVMs.

If a nonlinear kernel is used, because of the nonlinear mapping relation between the input space and the feature space, the linear discriminant function constructed by an SVM in the feature space corresponds to a nonlinear function in the original input space. The function family richness and discriminant power of SVMs are thus

incorporated by the mapping function and ultimately the kernel function, while problem formulation is kept in the same and neat form.

3.2 SVM-RFE

The SVM-RFE method was originally proposed to perform gene selection for cancer classification.^[72] Nested subsets of features are selected in a sequential backward elimination manner, which starts with all the features and removes one feature each time. At each step, the squares of the coefficients of the weight vector \mathbf{w} of a linear SVM are used as the feature ranking criterion.

Using w_i^2 as ranking score corresponds to removing the feature whose removal changes the objective function least. This objective function is chosen to be $J = \frac{1}{2} \|\mathbf{w}\|^2$ in SVM-RFE. This is explained by the optimal brain damage (OBD) algorithm,^[76] which approximates the change in objective function caused by removing the i th feature by expanding the objective function in Taylor series to second order (equation 7):

$$\Delta J(i) = \frac{\partial J}{\partial w_i} \Delta w_i + \frac{\partial^2 J}{\partial w_i^2} (\Delta w_i)^2 \tag{Eq. 7}$$

At the optimum of J , the first-order term can be neglected and with $J = \frac{1}{2} \|\mathbf{w}\|^2$, equation 6 becomes equation 8:

$$\Delta J(i) = (\Delta w_i)^2 \tag{Eq. 8}$$

where $(\Delta w_i)^2 = w_i^2$ corresponds to removing the i th feature.

The recursive elimination procedure of SVM-RFE is as follows:^[72]

1. Start: ranked feature $R = []$; selected subset $S = [1, \dots, d]$;
2. Repeat until all features are ranked:
 - (a) train a linear SVM with all the training data and variables in S
 - (b) compute the weight vector using equation 6
 - (c) compute the ranking scores for features in S : $c_i = (w_i)^2$
 - (d) find the feature with the smallest ranking score: $e = \arg \min_i c_i$
 - (e) update R : $R = [e, R]$
 - (f) update S : $S = S - [e]$;
3. Output: Ranked feature list R .

By eliminating one variable at each recursive step, eventually all the feature variables are ranked and a ranked feature variable list R is produced. The earlier one feature variable is eliminated, the lower it is ranked by SVM-RFE. For speed reasons, the algorithm can be generalized to remove more than one feature per

step. However, this speed-up strategy may degrade performance of SVM-RFE, especially if several feature variables are eliminated each time during the later stage of the recursive procedure.

Note that, in SVM-RFE, the following SVM formulation is used (equation 9):

$$\begin{aligned} \min_{\mathbf{w}, \mathbf{b}, \xi_i} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^l \xi_i^2 \\ \text{subject to} \quad & y_i(\mathbf{w} \cdot \mathbf{z}_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \end{aligned} \quad (\text{Eq. 9})$$

This formulation of SVM is usually solved by the following dual problem with a slightly modified kernel function $\tilde{k}(\cdot, \cdot)$ [equation 10]:

$$\begin{aligned} \min_{\alpha_i} \quad & \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j y_i y_j \tilde{k}(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^l \alpha_i \\ \text{subject to} \quad & \alpha_i \geq 0, \sum_{i=1}^l \alpha_i y_i = 0 \end{aligned} \quad (\text{Eq. 10})$$

where $\tilde{k}(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j) + 1/C \delta_{ij}$; $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise. Equation 9 uses a loss function different from equation

1 and the second term of the objective function in equation 1, $\sum_{i=1}^l \xi_i$,

is replaced by $\sum_{i=1}^l \xi_i^2$ in equation 9.

3.3 Numerical Experiments

We evaluate the SVM-RFE peak selection method together with the T-statistics method on two cancer classification MS datasets, lung cancer and ovarian cancer. The lung cancer dataset originally was provided for the First Annual Proteomics Datamining Conference, organized by the Department of Radiology and Biostatistics at Duke University in September 2002. Wagner et al.^[39] conducted some basic preprocessing steps on this dataset and reduced the number of peaks from 603 to 229. We used the preprocessed dataset (with 229 peaks) obtained from Wagner. We obtained the ovarian cancer dataset from Kent Ridge Bio-medical Data Set Repository.^[77] Neither dataset originally had a test set.

For performance validation, we retained some samples for testing purposes. Thus, we randomly split the original datasets into training and test sets and maintained the same percentages of the positive and negative samples in the training and test sets. In table I we summarize some basic information about the datasets, including the number of peaks, and the size of the training and test sets.

Table I. Number of peaks and sizes of training and test sets in the lung and ovarian cancer datasets

Dataset	Peaks	Training samples	Test samples
Lung cancer	229	29	12
Ovarian cancer	15 154	177	76

More detailed information about the two datasets can be found in the original reports.^[39,77]

In our study, we performed peak selection solely on the training set of each dataset. The goodness of a selected peak subset was evaluated by the performance of a classifier built on the training set, with only the selected set of peaks as input variables. We chose linear SVM as the classification algorithm, as linear methods are commonly used in cancer classification with MS data (for more information see Wagner et al.^[78]).

Classification error on test set (test error) is usually used to assess the performance of a classifier. However, the total number of available samples in our MS datasets is small. In such a case, the test error may be biased due to an 'unfortunate' partition of training and test sets. Thus, instead of reporting such a test error from one division of training and test sets, we merged the training set and test set and then partitioned the total samples again into a training and test set randomly by stratified sampling. We repeated the partition process 100 times. For each partition, we trained a linear SVM classifier on the training set (hyperparameter C is to be selected by 10-fold cross-validation on the training set) and then tested it on the corresponding test set. From these 100 trials we could compute the averages of performance measures.

To speed up the feature selection procedure of the SVM-RFE, when the number of features m is larger than the feature subset S selected at a time, we eliminated r ($r \geq 1$) features each time in our numerical experiments. We chose $r = 100$ if $m > 10\,000$, $r = 10$ if $1000 < m \leq 10\,000$ and $r = 1$ if $m \leq 1000$.

To compare our results, we used T-statistics for input feature selection, which selects the feature variables that individually are most relevant to the concept under study. A ranking score was computed for each feature, using the following feature ranking criterion (equation 11):

$$c_i = \frac{|\mu_i^+ - \mu_i^-|}{\sqrt{\frac{(\sigma_i^+)^2}{n^+} + \frac{(\sigma_i^-)^2}{n^-}}} \quad (\text{Eq. 11})$$

where μ_i^+ and μ_i^- are the mean values of the i th feature respectively over positive and negative samples; σ_i^+ and σ_i^- are the

corresponding standard deviations; n^+ and n^- denote the number of positive and negative training samples. Equation 11 fundamentally measures the normalized feature value difference between two sample groups.

For each feature subset selected by either the T-statistics method or SVM-RFE in our study, we computed the mean and standard deviation of the test error, sensitivity, and specificity, from 100 repetitions of training and testing. As we were mostly interested in small peak subsets, we evaluated the two methods only on peak subsets with the number of peaks ranging from 1 to 50. We plotted the average test error versus the size of feature subsets selected by two methods on the two datasets respectively in figures 2 and 3.

SVMs are capable of dealing with large numbers of input variables with little increase in computation complexity. To determine if feature selection improves the performance of SVMs, we also trained and tested SVMs with a full number of features on the same 100 partitions of training and test sets. The means and standard deviations of the test performance of SVMs with full features are reported in table II, together with those of the best feature subsets selected by T-statistics and SVM-RFE, with the number of selected peaks confined to less than 20.

4. Discussions and Conclusion

Ever since Petricoin et al.^[57] published their report on the diagnosis of ovarian cancer by using proteomic approaches, several groups have embarked on similar studies while re-analyzing and disputing their analysis results.^[31-33] However, some research-

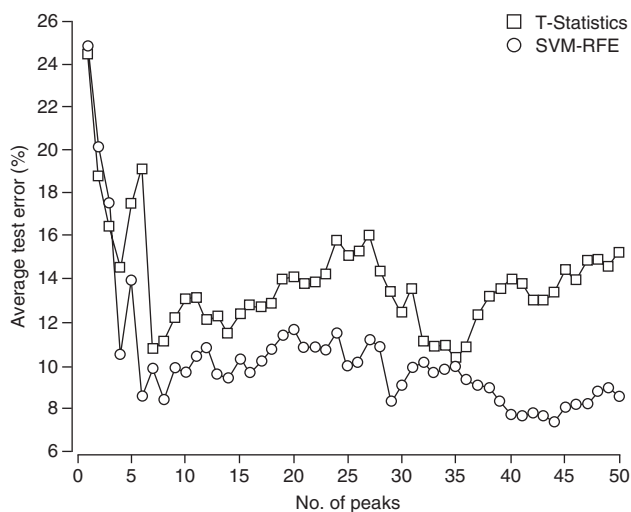


Fig. 2. Average test error rates at different sizes of peak subsets, selected by T-statistics and support vector machine-recursive feature elimination (SVM-RFE), on the lung cancer dataset.

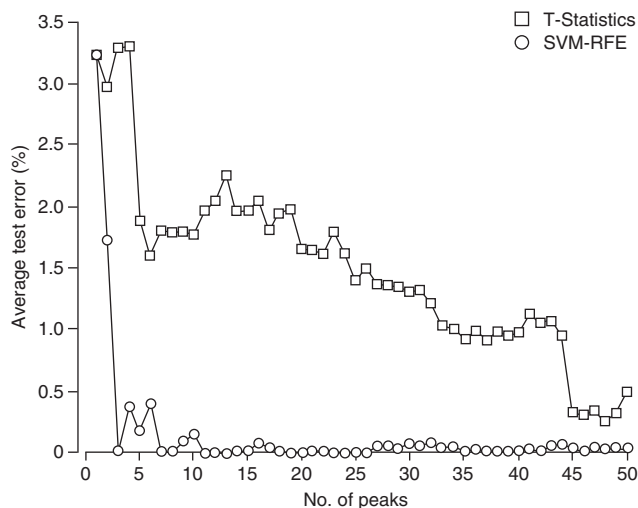


Fig. 3. Average test error rates at different sizes of peak subsets, selected by T-statistics and support vector machine-recursive feature elimination (SVM-RFE), on the ovarian cancer dataset.

ers^[79,80] believe that the development of protein profiling methods as diagnostic tools is still only in its early stages and consequently not ready for clinical use.^[34,81] In this review, we highlighted major processes and technologies involved in cancer classification by using MS data. For instance, we presented the importance of feature selection by demonstrating SVM-RFE method on lung and ovarian cancer data.

As illustrated, it is clear that SVM-RFE selects better peak subsets than the T-statistics feature selection method on the benchmark datasets. Higher classification accuracy was achieved with only a small number of peaks as input variables. Looking at the performance of SVMs without peak selection and SVMs with peak selection, we can see that the classification performance of SVMs with peak selection is much better than that of SVMs with all peaks as input variables. This observation tells us that selecting a subset of peaks not only improves the efficiency of classification algorithms but also improves the prediction accuracy, even for classification algorithm like SVMs, which can handle large numbers of input variables with little increase in computation complexity. Selecting a small number of peaks also avoids spurious results with MS data, for which the number of training samples is usually small compared with the number of peaks in the mass spectra. The high prediction accuracy also further strengthens our belief in the promising application prospects of MS patterns in the future cancer classification.

While it is known that T-statistics selects the peaks whose intensities differs most between the cancer and noncancer groups, the way that SVM-RFE selects the peak subset is not well understood. Checking the T-statistic scores of the peaks selected with

Table II. Performance of support vector machine (SVM) without peak selection and the performance of SVM with peak selection by T-statistics or SVM-recursive feature elimination (RFE), on the two datasets

	Number of peaks	Test error (%)	Sensitivity (%)	Specificity (%)
Lung cancer				
SVM	Full (229)	21.58 ± 9.63	90.29 ± 11.28	61.80 ± 21.29
T-statistics	7	10.75 ± 8.89	95.43 ± 7.56	80.60 ± 22.28
SVM-RFE	8	8.41 ± 5.98	94.57 ± 7.54	87.40 ± 13.23
Ovarian cancer				
SVM	Full (15,154)	0.50 ± 1.04	99.85 ± 0.52	98.85 ± 2.72
T-statistics	6	1.61 ± 1.39	99.31 ± 1.86	96.74 ± 2.70
SVM-RFE	11	0.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00

SVM-RFE may provide some insight into the way SVM-RFE works. With the lung cancer dataset, we found the T-statistic ranks of the eight peaks in the best subset selected by SVM-RFE were 36, 29, 17, 7, 4, 3, 2, and 1 (peak with rank 1 has the largest T-statistic score). On the ovarian cancer dataset, the T-statistic ranks of the 11 peaks in the best subset selected by SVM-RFE were 37, 50, 45, 14, 7, 6, 5, 4, 3, 2, and 1. For both datasets, the best peak subsets selected by SVM-RFE always contained peaks that were top ranked by T-statistics, while they also included some peaks not top ranked by T-statistics. However, the peaks selected by SVM-RFE together achieve much smaller test error than the same number of most top ranked peaks selected by T-statistics. Further investigation to identify the proteins underlying these selected peaks is needed in order to better understand the way SVM-RFE works and to gain insight into the disease pathway.

Acknowledgments

We wish to thank Michael Wagner at the Cincinnati Children's Hospital Medical Center for sharing his preprocessed lung cancer dataset with us.

No sources of funding were used to assist in the preparation of this paper. The authors have no conflicts of interest that are directly relevant to the content of this review.

References

- Seliger B, Kellner R. Design of proteome-based studies in combination with serology for the identification of biomarkers and novel targets. *Proteomics* 2002; 2 (12): 1641-51
- Banks RE, Dunn MJ, Hochstrasser DF, et al. Proteomics: new perspectives, new biomedical opportunities. *Lancet* 2000; 356 (9243): 1749-56
- Steiner S, Witzmann FA. Proteomics: applications and opportunities in preclinical drug development. *Electrophoresis* 2000; 21 (11): 2099-104
- Harry JL, Wilkins MR, Herbert BR, et al. Proteomics: capacity versus utility. *Electrophoresis* 2000; 21 (6): 1071-81
- Rai AJ, Zhang Z, Rosenzweig J, et al. Proteomic approaches to tumor marker discovery: identification of biomarkers for ovarian cancer. *Arch Pathol Lab Med* 2002; 126 (12): 1518-26
- Hanash SM, Bobek MR, Rickman DS, et al. Integrating cancer genomics and proteomics in the post-genome era. *Proteomics* 2002; 2 (1): 69-75
- Bichsel VE, Liotta LA, Petricoin EF. Cancer proteomics: from biomarker discovery to signal pathway profiling. *Cancer J* 2001; 7 (1): 69-78
- Rai AJ, Chan DW. *Cancer proteomics: serum diagnostics for tumor marker discovery*. New York: New York Academy of Sciences, 2004: 286-94
- Srinivas PR, Srivastava S, Hanash S, et al. Proteomics in early detection of cancer. *Clin Chem* 2001; 47 (10): 1901-11
- Stevens EV, Liotta LA, Kohn EC. Proteomic analysis for early detection of ovarian cancer: a realistic approach? *Int J Gynecol Cancer* 2003; 13: 133-9
- Plesch FN, Kubicka S, Manns MP. Prevention of hepatocellular carcinoma in chronic liver disease: molecular markers and clinical implications. *Dig Dis* 2001; 19 (4): 338-44
- Etzioni R, Urban N, Ramsey S, et al. The case for early detection. *Nat Rev Cancer* 2003; 3 (4): 243-52
- Shin BK, Wang H, Hanash S. Proteomics approaches to uncover the repertoire of circulating biomarkers for breast cancer. *J Mammary Gland Biol Neoplasia* 2002; 7 (4): 407-13
- Wulfkuhle JD, Liotta LA, Petricoin EF. Proteomic applications for the early detection of cancer. *Nat Rev Cancer* 2003; 3 (4): 267-75
- Petricoin EF, Ornstein DK, Pawletz CP, et al. Serum proteomic patterns for detection of prostate cancer. *J Natl Cancer Inst* 2002; 94 (20): 1576-8
- Vlahou A, Giannopoulos A, Gregory BW, et al. Protein profiling in urine for the diagnosis of bladder cancer. *Clin Chem* 2004; 50 (8): 1438-41
- Wagner PD, Verma M, Srivastava S. Challenges for biomarkers in cancer detection. *Ann N Y Acad Sci* 2004; 1022 (1): 9-16
- Vitorino R, Lobo MJC, Ferrer-Correia AJ, et al. Identification of human whole saliva protein components using proteomics. *Proteomics* 2004; 4 (4): 1109-15
- Srivastava S, Gopal-Srivastava R. Biomarkers in cancer screening: a public health perspective. *J Nutr* 2002; 132 (8): 2471S-5S
- Chace DH. Mass spectrometry in the clinical laboratory. *Chem Rev* 2001; 101 (2): 445-77
- Bergman AC, Benjamin T, Alaiya A, et al. Identification of gel-separated tumor marker proteins by mass spectrometry. *Electrophoresis* 2000 Feb; 21(3): 679-86
- Issaq HJ, Conrads TP, Prieto DA, et al. SELDI-TOF MS for diagnostic proteomics. *Anal Chem* 2003; 75 (7): 148A-55A
- Lin ZS, Jensen SD, Lim MS, et al. Application of SELDI-TOF mass spectrometry for the identification of differentially expressed proteins in transformed follicular lymphoma. *Mod Pathol* 2004; 17 (6): 670-8
- Bhattacharyya S, Siegel ER, Petersen GM, et al. Diagnosis of pancreatic cancer using serum proteomic profiling. *Neoplasia* 2004 Sep-Oct; 6 (5): 674-86
- Wilson LL, Tran L, Morton DL, et al. Detection of differentially expressed proteins in early-stage melanoma patients using SELDI-TOF mass spectrometry. *Ann N Y Acad Sci* 2004; 1022: 317-22
- Bhattacharyya S, Epstein J, Suva LJ. Preliminary characterization of serum biomarkers for multiple myeloma with bone metastasis using SELDI-TOF mass spectrometry. 25th Annual Meeting of the American Society for Bone and Mineral Research; 2003 Sep 19-23; Minneapolis (MN). *J Bone Miner Res* 2003; 18 Suppl. 2: S310

27. Gretzer MB, Chan DW, Rootselaar CL, et al. Proteomic analysis of dunning prostate cancer cell lines with variable metastatic potential using SELDI-TOF. *Prostate* 2004; 60 (4): 325-31
28. Becker S, Cazares LH, Watson P, et al. Surface-enhanced laser desorption/ionization time-of-flight (SELDI-TOF) differentiation of serum protein profiles of BRCA-1 and sporadic breast cancer. *Ann Surg Oncol* 2004; 11 (10): 907-14
29. Barker J. *Mass spectrometry*. 2nd ed. ACOL series. Chichester: John Wiley & Sons, 1998
30. Tang N, Tornatore P, Weinberger SR. Current developments in SELDI affinity technology. *Mass Spectrom Rev* 2004; 23 (1): 34-44
31. Baggerly KA, Morris JS, Coombes KR. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics* 2004; 20 (5): 777-85
32. Baggerly KA, Morris JS, Edmonson SR, et al. Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer. *J Natl Cancer Inst* 2005 Feb; 97 (4): 307-9
33. Sorace JM, Zhan M. A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics* 2003 Jun; 4: 24
34. Semmes OJ, Feng Z, Adam BL, et al. Evaluation of serum protein profiling by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry for the detection of prostate cancer: I. Assessment of platform reproducibility. *Clin Chem* 2005; 51 (1): 102-12
35. Seibert V, Wiesner A, Buschmann T, et al. Surface-enhanced laser desorption ionization time-of-flight mass spectrometry (SELDI TOF-MS) and ProteinChip® technology in proteomics research. *Pathol Res Pract* 2004; 200 (2): 83-94
36. Diamandis EP. Mass Spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. *Mol Cell Proteomics* 2004; 3 (4): 367-78
37. Fenn JB, Mann M, Meng CK, et al. Electrospray ionization for mass-spectrometry of large biomolecules. *Science* 1989; 246 (4926): 64-71
38. Fenn JB. Electrospray wings for molecular elephants (Nobel lecture). *Angew Chem Int Ed Engl* 2003; 42 (33): 3871-94
39. Wagner M, Naik D, Pothen A. Protocols for disease classification from mass spectrometry data. *Proteomics* 2003; 3 (9): 1692-8
40. Barclay VJ, Bonner RF, Hamilton IP. Application of wavelet transforms to experimental spectra: smoothing, denoising, and data set compression. *Anal Chem* 1997; 69 (1): 78-90
41. Sauve AC, Speed TP. Normalization, baseline correction and alignment of high-throughput mass spectrometry data. *Proceedings of the Workshop on Genomic Signal Processing and Statistics (GENSIPS 2004)*; 2004 May 26-27; Baltimore (MD)
42. Malmquist G, Danielsson R. Alignment of chromatographic profiles for principal component analysis: a prerequisite for fingerprinting methods. *J Chromatogr A* 1994; 687 (1): 71-88
43. Bylund D, Danielsson R, Malmquist G, et al. Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography-mass spectrometry data. *J Chromatogr A* 2002; 961 (2): 237-44
44. Gong F, Liang YZ, Fung YS, et al. Correction of retention time shifts for chromatographic fingerprints of herbal medicines. *J Chromatogr A* 2004; 1029 (1-2): 173-83
45. Johnson KJ, Wright BW, Jarman KH, et al. High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis. *J Chromatogr A* 2003; 996 (1-2): 141-55
46. Tomasi G, van den Berg F, Andersson C. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *J Chemometrics* 2004; 18 (5): 231-41
47. Pravdova V, Walczak B, Massart DL. A comparison of two algorithms for warping of analytical signals. *Anal Chim Acta* 2002; 456 (1): 77-92
48. Walczak B, Wu W. Fuzzy warping of chromatograms. *Chemometrics Intelligent Lab Systems* 2005; 77: 173-80
49. Eilers PHC. Parametric time warping. *Anal Chem* 2004; 76 (2): 404-11
50. Wang CP, Isenhour TL. Time-warping algorithm applied to chromatographic peak matching gas chromatography/Fourier transform infrared/mass spectrometry. *Anal Chem* 1987; 59 (4): 649-54
51. Stoyanova R, Nicholls AW, Nicholson JK, et al. Automatic alignment of individual peaks in large high-resolution spectral data sets. *J Magn Reson* 2004; 170 (2): 329-35
52. Torgrip R, Aberg M, Karlberg B, et al. Peak alignment using reduced set mapping. *J Chemometrics* 2003; 17 (11): 573-82
53. Hansen ME, Smedsgaard J. A new matching algorithm for high resolution mass spectra. *J Am Soc Mass Spectrom* 2004; 15 (8): 1173-80
54. Yasui Y, Pepe M, Thompson ML, et al. A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics* 2003; 4 (3): 449-63
55. Kohavi R, George HJ. Wrappers for feature subset selection. *Artificial Intelligence* 1997; 97 (1-2): 273-324
56. Blum A, Langley P. Selection of relevant features and examples in machine learning. *Artificial Intelligence* 1997; (1-2): 245-71
57. Petricoin EF, Ardekani AM, Hitt BA, et al. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 2002; 359: 572-7
58. Li J, Zhang Z, Rosenzweig J, et al. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clin Chem* 2002; 48: 1296-304
59. Adam BL, Qu Y, Davis JW, 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
60. Lu B, Chen T. A suffix tree approach to the interpretation of tandem mass spectra: applications to peptides of non-specific digestion and post-translational modifications. *Bioinformatics* 2003 Sep; 19 Suppl. 2: ii113-21
61. Chen T, Kao MY, Tepel M, et al. A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *J Comput Biol* 2001; 8 (3): 325-37
62. Lu BW, Chen T. A suboptimal algorithm for de novo peptide sequencing via tandem mass spectrometry. *J Comput Biol* 2003; 10 (1): 1-12
63. Yan B, Pan C, Olman VN, et al. A graph-theoretic approach for the separation of b and y ions in tandem mass spectra. *Bioinformatics* 2005; 21 (5): 563-74
64. Frank A, Pevzner P. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem* 2005; 77 (4): 964-73
65. Ma B, Zhang KZ, Hendrie C, et al. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* 2003; 17 (20): 2337-42
66. Taylor JA, Johnson RS. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* 1997; 11 (9): 1067-75
67. Johnson RS, Taylor JA. Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry. *Mol Biotechnol* 2002; 22 (3): 301-15
68. Eng J, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 1994; 5: 976
69. Perkins DN, Pappin DJC, Creasy DM, et al. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999; 20 (18): 3551-67
70. Pappin DJ, Hojrup P, Bleasby AJ. Rapid identification of proteins by peptide-mass fingerprinting. *Curr Biol* 1993; 3 (6): 327-32
71. Zhang WZ, Chait BT. Profound: an expert system for protein identification using mass spectrometric peptide mapping information. *Anal Chem* 2000; 72 (11): 2482-9
72. Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines. *Machine Learning* 2002; 46 (1-3): 389-422
73. Boser B, Guyon I, Vapnik V. A training algorithm for optimal margin classifiers. In: Haussler D, editor. *Proceedings of the 5th Annual Workshop on Computational Learning Theory*; 1992: 114-52
74. Vapnik VN. *Statistical learning theory*. New York: John Wiley & Sons, 1998
75. Duan K-B, Rajapakse JC. SVM-RFE peak selection for cancer classification with mass spectrometry data. In: Chen P, Wong L, editors. *Advances in bioinformat-*

- ics and computational biology 1. Proceedings of the 3rd Asia-Pacific Bioinformatics Conference; 2005 Jan 17-21; Singapore. London: Imperial College Press, 2005: 191-200
76. LeCun Y, Denker J, Solla S, et al. Optimal brain damage. In: Touretzky DS, editor. *Advances in neural information processing systems II*. San Mateo (CA): Morgan Kaufmann, 1990
77. Li J, Liu H. Kent ridge bio-medical data set repository, 2002 [online]. Available from URL: <http://sdmc.lit.org.sg/GEDatasets/Datasets.html> [Accessed 2005 Jun 23]
78. Wagner M, Naik DN, Pothen A, et al. Computational protein biomarker prediction: a case study for prostate cancer. *BMC Bioinformatics* 2004; 5: 26
79. Pritzker KPH. Cancer biomarkers: easier said than done. *Clin Chem* 2002 Aug; 48 (8): 1147-50
80. Check E. Proteomics and cancer: running before we can walk? *Nature* 2004; 429 (6991): 496-7
81. Hortin GL. Can mass spectrometric protein profiling meet desired standards of clinical laboratory practice? *Clin Chem* 2005; 51 (1): 3-5
-
- Correspondence and offprints: Dr *Jagath C. Rajapakse*, School of Computer Engineering, BioInformatics Research Center, Nanyang Technological University, Block N4, 50 Nanyang Avenue, 639798, Singapore.
E-mail: asjagath@ntu.edu.sg