Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Biomedical Informatics 39 (2006) 227–248
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Methodological Review

A machine learning perspective on the development of clinical decision support systems utilizing mass spectra of blood samples

Hyunjin Shin [a], Mia K. Markey [b,*]

[a] Electrical and Computer Engineering Department, The University of Texas at Austin, USA
[b] Biomedical Engineering Department, The University of Texas at Austin, USA

Received 18 December 2004
Available online 26 May 2005

Abstract

Currently, the best way to reduce the mortality of cancer is to detect and treat it in the earliest stages. Technological advances in genomics and proteomics have opened a new realm of methods for early detection that show potential to overcome the drawbacks of current strategies. In particular, pattern analysis of mass spectra of blood samples has attracted attention as an approach to early detection of cancer. Mass spectrometry provides rapid and precise measurements of the sizes and relative abundances of the proteins present in a complex biological/chemical mixture. This article presents a review of the development of clinical decision support systems using mass spectrometry from a machine learning perspective. The literature is reviewed in an explicit machine learning framework, the components of which are preprocessing, feature extraction, feature selection, classifier training, and evaluation.

Keywords: Diagnosis; Computer-assisted; Spectrum analysis; Mass spectrometry; Neoplasms; Blood; Artificial intelligence; Signal processing; Automatic data processing; Pattern recognition; Classification

1. Background and motivation

Cancer is a major public health concern in the US. In 2004, there will be more than 1.3 million new cancer cases and more than 563,000 deaths due to cancer [1,2]. Cancer accounts for one of every four deaths in the US [2]. Currently, the best way of reducing the mortality of cancer is to detect and treat it in the earliest stages [3]. For example, when breast cancer is detected at the advanced stage, in which cancer is metastasized from the original organ to others, the survival rate is only 23%. However, when breast cancer is detected at the early stage, in which cancer is localized in organ of origin, the survival rate increases to 97% [2]. Similarly, the survival rate of prostate cancer soars from 34% when the cancer is detected at the advanced stage to nearly 100% at the early stage [2].

A cancer screening test is considered efficacious if it results in a decrease in cause-specific mortality. Necessary evidence in favor of a particular screening test includes earlier detection of disease than would have occurred due to presentation of symptoms and evidence that earlier treatment will result in a better outcome. (There is a helpful overview online at http://cancer.gov/cancertopics/pdq/screening/overview.) Screening and diagnostic tests are typically evaluated in terms of their sensitivity and specificity. Sensitivity is the fraction of disease cases that are correctly identified as disease. Specificity is the fraction of non-disease cases that are correctly identified as non-disease.

Currently, there exist effective screening tests for use in the general population for only a few types of cancer. The screening methods that are best supported by the evidence to date are (1) the Pap smear for cervical cancer screening, (2) mammography for breast cancer detec-

* Corresponding author. Fax: +1 512 471 0616.
E-mail address: mia.markey@mail.utexas.edu (M.K. Markey).

1532-0464/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2005.04.002

tion, and (3) fecal occult blood testing for colorectal cancer screening. While there are limitations to each of these methods, there is evidence that they have made substantial contributions to reducing the morbidity and mortality due to cancer.

A Pap smear is an exfoliative cytological staining procedure that can help identify premalignant and malignant changes in the cervical epithelium. The incidence of, and mortality of women due to, cervical cancer has declined about 70% in the US since the Pap was introduced in the 1950s. Use of this screening test reduces the incidence as well as mortality since the Pap smear can detect precancerous changes that can be treated. However, with a specificity of only 63%, many false-positive Pap smears occur in screening the general population, in which cervical cancers and precancerous lesions are thankfully rare [4]. Unfortunately, false negative Pap smears also occur since the sensitivity of the exam is 73% [4].

Mammography, X-ray imaging of the breasts, is used to detect breast cancer. Mammography has reduced the mortality of breast cancer by approximately 25–30% in the US since the 1970s [5,6]. Mammography also suffers from false positives due to the combination of moderate specificity and low disease prevalence. Only 15–34% of the positive cases from mammography are found to be actually malignant at biopsy [7,8]. False negatives also occur since the sensitivity of mammography is approximately 90% [9].

Fecal occult blood testing (FOBT) is used for the early detection of colorectal cancer. It can detect colorectal cancer by measuring blood loss in the stool, which mainly occurs due to colorectal neoplasms [10,11]. FOBT is reported to have reduced the mortality of colorectal cancer in the US by 33% [12,13,10,11]. FOBT has a fairly high specificity of 96–98% [12]. However, because the sensitivity of the FOBT is merely 40% [12], there is concern that the diagnosis and treatment of colorectal cancer can be delayed due to false negative tests.

An ideal cancer screening method would be accurate, non-invasive, and inexpensive. As discussed above, the accuracy levels of existing screening methods are far from ideal. The false negatives resulting from screening methods in current use delay the diagnosis of cancer, which can lead to increased morbidity and mortality. The false positives generated by the early detection methods used in current practice necessitate additional diagnostic testing which increases costs, discomfort, and stress. Existing screening modalities are all invasive to some extent: a Pap smear is obtained from a pelvic exam, mammography is based on exposure to ionizing radiation and compression of the breasts, and FOBT requires a stool sample. Many variables are believed to impact compliance with existing screening programs, but physical discomfort and embarrassment are probably important factors (e.g., [14]). The costs associated with current approaches to cancer screening remain problematic as well (e.g., [15]).

Recent technological advances in genomics and proteomics have opened a new realm of early detection, showing potential to overcome the drawbacks of current early detection strategies. A biomarker is a biologically derived molecule in the body that indicates the progress or status of a disease. The concentration level or pattern of biomarkers related to a certain type of cancer can be used for early detection or diagnosis. Studies of the application of biomarkers for early cancer detection can be summarized into two categories: the usage of a single biomarker and the pattern analysis of multiple biomarkers.

When a single biomarker is used, the concentration level of the biomarker is taken as an indicator of the presence or absence of cancer. A threshold is set on the concentration level of a biomarker and if the concentration level is higher than the threshold, the specimen is considered ''positive'' for cancer. An example of early detection based on a single biomarker is the use of prostate-specific antigen (PSA) in blood to detect prostate cancer. PSA is a protein secreted by the epithelial cells of the prostate gland. The PSA level in blood is generally low in healthy people or patients with benign prostate disease such as benign prostatic hyperplasia (BPH), but it tends to rise in many patients with malignancies [16]. However, the specificity of using the concentration level of PSA as an indicator of prostate cancer ranges from only 18 to 50% with a sensitivity of 70–90% [16]. The low specificity causes many false positives to occur; therefore, unnecessary biopsies are performed to corroborate the absence of prostate cancer. There is considerable debate as to whether screening for prostate cancer by PSA is efficacious [3,17–20].

The problems encountered with the PSA biomarker suggest limitations that may plague any test based on a single biomarker. Given the high level of biological variability and the fact that cancer cells are derived from normal cells in the body, it may not be possible to identify a single circulating protein that can identify the presence of cancer with high sensitivity and specificity in the general population. Even for high-risk populations (e.g., CA 125 for women at high risk for ovarian cancer [21]), it is unlikely that single biomarkers will provide as accurate testing as the use of multiple biomarkers.

An important difficulty in developing tests based on single biomarkers is that the identification process demands a vast amount of time and labor [20]. Traditionally, 2D gel electrophoresis (2DE) has been used for biomarker identification in tandem with mass spectrometry [22,23,19,24,20]. A protein expressed differently between cancer and normal specimens is extracted using 2DE and the extracted protein is identified by peptide fingerprinting using mass spectrometry and protein/

peptide databases. 2DE is the bottleneck of this process because it is extremely time-consuming and laborious [25,26,20].

Recently, pattern analysis of multiple biomarkers in blood samples has attracted attention as an alternative to the usage of a single biomarker for early detection of cancer. The pattern differences of protein profiles between cancer and healthy samples are perceived using data mining algorithms. Multiple proteins rather than a single protein are used as a 'panel' of biomarkers in this approach. Because these pattern differences originate from the complexity of blood, which is a mixture of thousands of proteins, a protein profiling modality with high-throughput and high sensitivity is required.

Mass spectrometry has the potential to meet these requirements by providing the sizes and relative abundances of the proteins in a complex biological/chemical mixture in a rapid and precise manner [27–31]. Recently, studies have been performed on a several types of cancer, including ovarian [32–43], prostate [44–50,40, 51–53], breast [54,55], bladder [56,57], lung [58–72], liver [73], pancreatic [74–79], renal cell carcinoma [80], colorectal [81], and astroglial tumor [82]. Most of these studies reported fairly high sensitivities and specificities (over 80%). However, many questions have been raised about the reliability of these reported results due to the ''black box'' methods employed [83–86,53,87].

Diamandis [85] pointed out that the peak height does not linearly correspond to the protein abundance because mass spectrometry only provides the *relative* abundance of proteins in a sample. He also inquired about why different data mining algorithms had produced different sets of potential biomarkers. He took as an example the studies on prostate cancer performed by Qu et al. [49] and Petricoin et al. [88]. They achieved high sensitivities (96%: Qu et al.; 95%: Petricoin et al.) and specificities (98%: Qu et al.; 83%: Petricoin et al.) with different sets of potential biomarkers selected through different data mining algorithms. Another question is why known biomarkers, for example PSA, do not seem to be reflected by the studies so far and the potential biomarkers found in these studies have fairly low mass [84–86]. Since low mass proteins are easily cleared by the kidney, the efficacy of a panel of low mass proteins appears to be suspicious. Related to this, Diamandis and Merwe [87] also raised another question on whether or not the putative biomarkers identified through the ''black box'' methods originate from cancer-specific pathological states in the body. They took an example Koomen et al.'s [75] study on the identification of potential biomarkers for pancreatic cancer. Koomen et al. [75] identified several biomarker candidates for pancreatic cancer from mass spectra of human plasma of healthy people and pancreatic cancer patients using statistical and biochemical tests. However, Diamandis and Merwe argued that these biomarker candidates can be only high abundance non-cancer-specific proteins in blood, which are produced by non-specific epiphenomena of cancer presence. Moreover, they suspected that the current mass spectrometers such as MALDI-TOF or SELDI-TOF are not sensitive enough to detect low abundance clinically useful biomolecules without an aid of powerful fractionation [87].

In addition to Diamandis' questions, Baggerly et al. [83,89] emphasized the problems in quality control indicated by the lack of reproducibility of the studies of Petricoin et al. [88] and Zhu et al. [43]. In these both studies, the ovarian cancer data sets posted on the website of the clinical proteomics program under the national cancer institute (http://home.ccr.cancer.gov/ncifdaproteomics/) were analyzed to identify diagnostic signatures for ovarian cancer. Baggerly et al. attempted to reproduce the experimental results obtained by Petricoin et al. by following the proposed bioinformatic algorithms as much as possible; however, Baggerly et al.'s [83] analyses imply that the apparent successes of the study may have been due to artifacts of sample processing rather than actual biological pattern differences. In the analyses on Zhu et al.'s study, Baggerly et al. [89] also showed that the peaks identified as potential biomarkers in one data set may not have consistently occurred in another set measured on a different date from the first set. Similarly, Yasui et al. [53] discussed the variability of the relative abundance of the same protein across chips and samples, which also points to the need for active and systematic internal quality controls.

Recently, some progress has been made in addressing these important questions. For example, low mass biomarkers may be more meaningful than many had believed because other high abundance and high mass proteins such as albumin can act as carriers of low mass biomarkers. These carrier proteins enable low mass biomarkers to stay in the body longer than expected [90]. Powerful fractionation techniques amplify the concentration of these low mass biomarkers by isolating them from the carrier proteins such that mass spectrometers can sufficiently detect the pathological signatures of these low mass biomarkers [17,90].

In addition, in response to Baggerly et al. [89], Liotta et al. pointed out that those two ovarian data sets used in Zhu et al.'s study were measured under different experimental settings (e.g., chemistries on protein chips, pH, laser energy intensity, etc.) as well as on different days; thus, simple comparisons of two different mass spectral data reproducibility may lead to a hasty generalization [91]. Similarly, Grizzle et al. [92] also maintained that it would be very unlikely for different laboratories to derive similar sets of biomarker candidates when applying different bioinformatics algorithms to samples obtained from non-identical patient populations.

Such issues related to reproducibility can be resolved to some extent if strongly standardized calibration and

instrumentation protocols are shared among laboratories. Recently, Semmes et al. [93] reported that "between-laboratory" reproducibility of SELDI-TOF MS can reach "within-laboratory" reproducibility levels if calibration and instrumentation protocols are strongly standardized among different laboratories. Six different institutions succeeded in classifying prostate cancer sampled from healthy samples using a classifier trained in an institution within an acceptable variance of error rates after calibrating the SELDI-TOF MS machines with the standard pooled serum samples distributed by one of these institutions. This study was performed as a part of an on-going effort to validate the approach of cancer detection through serum protein expression profiling using SELDI-TOF MS [94].

However, many questions remain unanswered. In Semmeset al.'s study, while mass accuracy of the healthy samples used for the quality control agreed within an acceptable variance, their peak intensities, especially small peak intensities, showed fairly high variation despite of careful calibration. Moreover, for classification, Semmes et al. selected prostate cancer and healthy samples that had been used in building the classifier in their previous study and on which the classifier performed well. Thus, as Semmes et al. discuss in their article, their study only shows the possibility that the experimental platform can be reproducible under very rigorous unified calibration and instrumentation protocols and more work is needed on this important issue.

For reliable early detection based on pattern analysis of multiple biomarkers, more rigorous and systemic approaches are needed. In this article, we review the literature on the development of clinical decision support systems using mass spectrometry in an organized framework from a machine learning perspective. Study design and quality control (e.g., sample preparation and mass spectrometer parameter settings) are also extremely important issues because data quality, which is mostly determined by these processes, affects the overall performance of decision support systems. However, since these issues are beyond the scope of this article, we will refer the reader to other papers that have discussed the topic of study design and quality control for experiments based on protein profiling techniques [95–98,93].

2. Mass spectrometry

Mass spectrometry provides rapid and precise measurements of the sizes and relative abundances of the proteins present in a complex biological/chemical mixture. Here we provide a very brief overview of the technique as it is typically used for identifying cancer biomarkers from blood samples. We refer the reader to other articles for a thorough review of mass spectrometry methods [99–104].

The capabilities of a mass spectrometer are determined by its ion source, mass analyzer, and detector. Protein profiling of plasma and serum has been performed primarily with a matrix-assisted laser desorption ionization (MALDI) ion source or its derivative, the surface-enhanced laser desorption ionization (SELDI) ion source coupled to a time-of-flight (TOF) mass analyzer with a chevron microchannel plate detector. The only difference between SELDI and MALDI is the use of derivatized surfaces to capture peptides and proteins based on particular physical or biochemical characteristics prior to MALDI sample preparation and mass analysis. A brief description of MALDI-TOF mass analysis is given in the following paragraphs.

To prepare proteins or peptides for MALDI mass analysis, aqueous solutions of the proteins or peptides are mixed with solutions of matrix molecules, like sinapinic acid and α-cyano-4-hydroxycinnamic acid, which are present in large molar excess compared to the proteins and peptides (10,000:1). Aliquots of this mixture are deposited on the MALDI plate and allowed to dry (this procedure is referred to as the dried droplet technique). The peptides and proteins selectively cocrystallize with the MALDI matrix as the solvent evaporates. After drying, the sample plate is introduced into the vacuum chamber of the mass spectrometer and placed in the MALDI ion source. To produce ions, an ultraviolet laser (337 or 355 nm) is used to irradiate the matrix crystals. The energy from these photons is transferred into translational and vibrational energy causing desorption of matrix material containing the peptide and protein analytes. The softer process ionization of MALDI (when compared to laser desorption ionization) prevents fragmentation of the protein and peptide analytes [103]. However, ionized clusters of matrix molecules produce chemical noise, which interferes with the ion signals of interest: those corresponding to the peptides and proteins [105,106].

After a delay of a few hundred nanoseconds (Wiley–McLaren time lag focusing), all ions are extracted from the source and accelerated into the TOF mass analyzer. The voltage settings in the ion source determine the range of optimized ion signal; i.e., the TOF has mass-dependent focusing. The ions drift in a field free region, where they are separated based on their mass-to-charge ratios. The principle behind this separation is that the potential energy of each ion in an electric field ($U = zV$) is converted into the kinetic energy of the ion in the TOF ($E = \frac{1}{2}mv^2$). By setting these equations equal to one another, the TOF equation can be derived and rearranged to calculate m/z value for an ion:

$$zV = \frac{1}{2}mv^2, \tag{1}$$

$$v = \frac{l}{t}, \tag{2}$$

$$\frac{(2Vt^2)}{l^2} = \frac{m}{z}. \quad (3)$$

In Eq. (1), z denotes the ion's charge amount, V is the electric potential that accelerates the ion, m is the ion's mass, and v is the ion's velocity. In Eq. (2), l is the length of the flight tube of the TOF mass spectrometer and t the flight time of the ion. Eq. (3) shows that the mass-to-charge ratio can be represented as a quadratic function of the flight time. Ions of the same m/z have the same flight time and thus impact the detector at the same time. When the ion strikes the detector, a cascade of secondary electrons is released. This current is captured by an anode and converted to a voltage using a preamplifier. The resulting voltage is recorded by a digital storage oscilloscope or by a digitizer card in a computer, and the amplitude of the signal corresponds to the number of ions that struck the detector in each bin of ion flight time. Other sources of noise from physical and electrical components of the mass spectrometer are also recorded (e.g., high frequency noise).

Data are recorded as plots of intensity versus flight time and displayed as intensity versus m/z: referred to as a mass spectrum (Fig. 1). In each mass spectrum, the individual ion signals correspond to non-volatile analytes in the original sample. In protein profiling, these ion signals primarily correspond to peptides and proteins because of the analyte specificity of the matrices described above. The mass-to-charge ratios (m/z), displayed as the x-axis, can be used to calculate the molecular weights of protein or peptide in the profile. For the analysis of complex mixtures, like plasma or serum protein fractions, MALDI-TOF MS has detection sensitivity in the 0.1–10 pmol range and mass measurement accuracy ranging from 0.01 to 0.5%. Ion signals in different mass spectra with centroids m/z values within the mass measurement error tolerance should be considered to be the same peak (protein). Because of the complexity of the samples, which produces suppression effects, and the lack of internal and external standards for quantification, the intensity of the ion signals in the protein profiles does not directly correlated to protein concentration [90]. Nonetheless, relative abundances of a particular ion signal can be determined by comparing mass spectra acquired from different samples. Thus, noise reduction and normalization schemes are critical to enable accurate statistical analysis of mass spectra.

Ciphergen (Freemont, CA) developed a SELDI-TOF system to accomplish both fractionation and mass analysis in a succinct and accurate manner. SELDI-TOF is a special case of MALDI-TOF in which chromatography is performed using protein chips that can capture only those proteins that biochemically/chemically match certain binding characteristics (e.g., hydrophobic), even when a variety of proteins are mixed together in high concentrations [107–109,31,110,111]. The selected "fraction" of proteins deposited on the protein chip is analyzed through MALDI-TOF mass spectrometry. SELDI-TOF enables to amplify the mass abundance information on more proteins than other types of mass spectrometry using the protein chip with the predefined chromatographic surface [90].

Recently, more advanced types of mass spectrometery have been tested to improve the sensitivity to diagnostic patterns in protein profiling [33,35,48,51]. Whereas the traditional mass spectrometers provide 15,000–40,000 m/z data records, high-resolution mass spectrometers can extend these to 350,000–400,000 [35]. The hybrid quadrupole time-of-flight (QqTOF) such as QSTAR pulsar I (Applied Biosystems, Framingham, MA, USA) is a frequently used model for this purpose. To the best of our knowledge, there have been no studies in which types high-resolution mass spectrometers were extensively compared and discussed. One expects that the development of more efficient and effective preprocessing and feature extraction/selection algorithms will be even more important issues for high-resolution MS than in traditional MALDI-TOF or SELDI-TOF because of the increase in size of each of data record.

3. Blood samples

This article reviews approaches that are being explored for cancer diagnosis using mass spectrometry of blood samples. There are several advantages to using

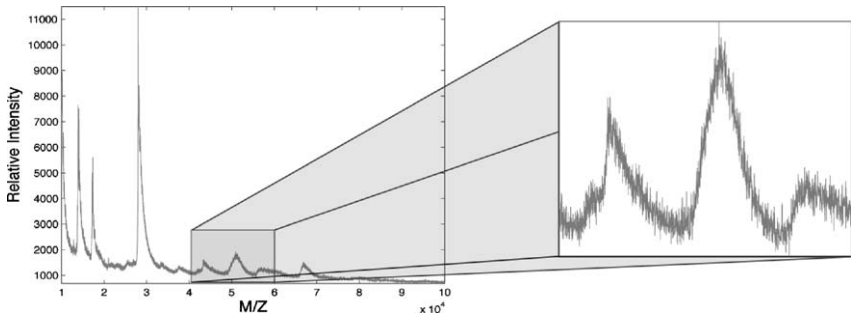

Fig. 1. Example of a mass spectrum in which the relative abundance is plotted as a function of the mass-to-charge ratio (m/z). Notice the monotonically decreasing baseline. A portion of the spectrum has been enlarged so that the high frequency noise is apparent.

blood samples because blood is readily accessible. While more invasive than some diagnostic imaging modalities, there is relatively little discomfort and low risk of side effects or adverse events associated with blood testing. Obtaining blood samples is less expensive than many other procedures. The primary disadvantage of using blood samples is that one expects that tumors located in most organs of the body will produce few proteins that will circulate in the blood at an appreciable level.

Throughout our discussion, we refer generically to "blood samples"; however, the reader should note that mass spectrometry is not performed on whole blood but on derived products, particularly plasma or serum. Plasma is the liquid portion of blood in which the cells are suspended; serum is the fluid that remains after clotting proteins are removed from plasma [112]. The advantage of using plasma rather than serum is that it contains more proteins and that the protease activity, which leads to protein degradation, is inhibited in plasma but not in serum. However, the disadvantage of using plasma is that low abundance proteins associated with disease may be difficult to detect in the presence of a large amounts of common proteins involved in clotting. Both plasma and serum have been used in studies of cancer diagnosis using mass spectrometry and it is not yet known which is best for this kind of analysis.

There have been many studies of the serum/plasma proteomes using techniques such as 2D gel electrophoresis (e.g. [113]). However, to the best of our knowledge, for the most part this information has not been incorporated into studies of cancer diagnosis using mass spectrometry. It is possible that more accurate models for sample classification could be developed if prior knowledge of blood proteins could be properly taken into account.

4. Framework for system development

We employ a machine learning framework to review the literature on the development of clinical decision support systems utilizing mass spectrometry of blood samples. There are five stages of data analysis in this framework. First, the spectra are preprocessed to reduce the contribution of noise and to normalize the spectra from different samples such that they are comparable. Second, features reflecting the pathological status of a sample are extracted from the mass spectra. Interpretable features, such as peaks corresponding to distinct protein species, are generally preferred. Third, highly discriminant features are selected to reduce the dimensionality of the data, which increases the likelihood of successful classification. Fourth, machine learning models are designed to distinguish cancer from normal samples based on the selected features. Fifth, the system is evaluated in terms of clinically relevant metrics such as sensitivity and specificity. Ideally, separate data sets should be used for each stage. However, in practice some form of data partitioning of a single data set, such as cross-validation or bootstrap sampling, is employed due to the difficulties of obtaining a large number of spectra. The five stages are mutually dependent and the best combination of methods to be used at each stage must be determined empirically.

4.1. Preprocessing

Biomedical data are notoriously complex and variable. The goal of preprocessing methods is to "clean up" the data such that machine learning algorithms will be able to tease out key information and correctly classify new samples based on a limited set of examples. In analyzing mass spectra of blood samples, the preprocessing stage includes two main tasks: noise reduction and normalization.

In mass spectrometry, the noise is the undesired interfering signal caused by sources unrelated to the biochemical nature of the sample being analyzed and the signal is the relative abundance of ions originating from the proteins in the sample. Many studies to date have not employed explicit noise reduction schemes other than basic noise reduction methods implemented on commercial mass spectrometers (e.g., the SELDI-TOF mass spectrometer from Ciphergen, Freemont, CA). However, some investigators have explored methods for reducing noise, particularly the baseline and high frequency noise [58,96,114,61,62,64,115,69,43,116].

Mass spectra exhibit a monotonically decreasing baseline (Fig. 1). As described above, it is necessary to add a matrix material to the sample of interest. However, it is possible for the matrix material to interact with itself as well as with the sample proteins. The baseline originates from small clusters of matrix material. Because the chances of cluster formation decrease with cluster size, the baseline diminishes monotonically as the mass-to-charge ratio increases [105,102]. The monotonically decreasing baseline can be regarded as low frequency noise because the baseline lies over a fairly long mass-to-charge ratio range [117]. Most studies that have employed a baseline reduction method have taken a two-step approach: baseline estimation followed by subtraction of the estimated baseline from the original mass spectrum.

A variety of approaches have been explored to estimate the baseline from mass spectra. Such approaches can be summarized into two major categories: heuristic or model-based. Heuristic approaches form non-parametric estimates of the baseline from a set of mass spectra. Model-based approaches build a mathematical model of the baseline based on the physics of the mass spectrometer and estimate the parameters of the model from a set of mass spectra. The baseline estimated by

either approach is then subtracted from the original spectrum. So far, there have been many more studies using heuristic approaches [58,96,75,62,64,69] than model-based approaches [116].

There have been several studies in which a heuristic approach was used to estimate and eliminate the baseline. A local average or minimum intensity within a moving window has been used as a local estimator of the baseline and the overall baseline is estimated by sliding the window over the mass spectrum [58]. Piecewise linear regression has been applied to the regions with a monotonically decreasing baseline [64,69]. The baseline has also been estimated by calculating the convex hull of the intensities of the proteins in a region [62]. All these algorithms seem to effectively estimate the underlying baseline, at least in some circumstances. However, the parameters of these algorithms, e.g., the width of the window in a piecewise linear regression model, have been determined in an ad hoc manner. For methods in which a sliding window or piecewise linear regression are employed for baseline elimination, the window size is a critical factor determining the overall performance. If the window size is too large, these methods may oversimplify the curvature of the baseline with a long straight line. If the window size is too small, they may produces an overly complex estimate of the baseline, which is very sensitive to high frequency noise.

There are no absolute standards for deciding which one among the heuristic baseline estimation algorithms is more effective than the others; each algorithm has its strengths and weaknesses. For example, choosing the minimum peak intensity within the sliding window as a local baseline estimator is superior to piecewise linear regression in terms of computation time. However, the latter method is expected to be relatively less sensitive to high frequency noise than the former one because a straight line with the minimum sum of errors between the line and peak intensities within the windows is calculated as a local estimator by linear regression. The convex hull is defined as the minimal convex set of given objects [118]. Thus, the convex hull of a mass spectrum is the piecewise straight lines connecting the local minima on the spectrum. This can be easily visualized by imagining a rubber band tightly stretched to encompassing the lower side of the mass spectrum. Since the convex hull is calculated based on the local minima, it may also suffer from the interference from high frequency noise.

To the best of our knowledge, there has only been one model-based approach reported in the literature to date [116]. Malyarenko et al. [116] used a model for the baseline in SELDI-TOF was developed using the phenomenon of charge accumulation that decays exponentially on the ion detector. Greater emphasis will likely be placed on model-based approaches in the future because they may be more effective with limited data sets since a priori knowledge is taken into account.

Mass spectra of blood samples also exhibit an additive high frequency noise component (Fig. 1). The presence of this noise hampers both data mining algorithms and human observers in finding meaningful patterns in mass spectra. While several prior studies have explored methods for reducing the influence of this high frequency noise [114,61,62,116,115,119,43], few have attempted to identify or describe the sources of this noise or to determine proper models for its statistical characteristics [120,96,105,106,117]. Moreover, to date, no study has used such noise characterization work to develop a ''model-based'' high frequency noise reduction scheme.

The heuristic high frequency noise reduction approaches employed most commonly in studies to date are smoothing filters [62,119,43], the wavelet transform (WT) [114,50,71], or the deconvolution filter [116]. Typical smoothing filters are the Gaussian filter [119,43] and moving average filter [62]. These smoothing filters smear out the high frequency noise signal in the spectra by averaging the intensities within a moving window. In the case of a Gaussian filter, the intensities are weighted by a Gaussian kernel before calculating the average. Over the past decade, the WT has been frequently used for chemical/biological signal processing [121,122]. The WT is a type of signal decomposition algorithm that allows us to view a signal as a superposition of weighted basis functions with different frequencies and time shifts. The frequency range and time location of the high frequency noise are localized using the WT. Then the high frequency noise can be effectively reduced by manipulating the weight coefficients of the basis functions [121,114,61,50,122]. The deconvolution filter reduces noise by minimizing the sum of squared errors between the desired output and filtered signal and the power of filtered noise. In this case, it is assumed that the observed signal can be modeled as the sum of the true signal and additive stationary noise [123]. Malyarenko et al. [116] applied the deconvolution filter to SELDI-TOF mass spectra and reported that it reduced noise and improved the resolution.

All of the methods have made considerable contributions to high frequency noise reduction in mass spectra. However, since no study has extensively compared the methods introduced above on the same data set, it is difficult to conclude if one method is better than the others. Moreover, the overall performance of those high frequency noise reduction methods is highly dependent on the choice of the filter parameters (e.g., the size of the sliding window or the kernel weights) and the true effectiveness of those methods is difficult to measure due to the lack of knowledge on the statistical characteristics of the signal and noise in mass spectra.

Most noise reduction approaches to date have emphasized designing filters based on empirical insight rather than rigorous statistical noise analysis. However,

a few studies have tried to identify the noise sources in mass spectrometry and to measure the statistical characteristics of the noise [120,96,105,106,117]. Such studies are critical because the lack of information on the statistical characteristics of the true signal and the noise may lead to the design of filters that remove the desired signal or fail to remove the noise. In other words, aggressive filtering may smear out diagnostically informative patterns and insufficient filtering may leave high levels of noise in the signal. Because low abundance proteins are expected to contain diagnostically useful information, noise reduction approaches that ignore statistical noise analysis may actually make it more difficult to detect differences in the spectral patterns between cancer and healthy samples. In future work, these noise characterization studies could provide the basis for model-based approaches to noise reduction.

A peak in mass spectra indicates the *relative* abundance of a protein; therefore, the magnitudes of mass spectra cannot be directly compared with each other. Normalization methods scale the intensities of mass spectra to make mass spectra comparable (Fig. 2). The most frequently used normalization method is normalization with respect to the total ion current (TIC), i.e., the sum of all the peaks in a mass spectrum [58,54,74,77,36,124,46,55,125,65,56]. Normalization with respect to the mean spectrum has also been used, which is equivalent to normalization with respect to TIC [43]. Other studies have performed normalization with respect to the largest peak [60,69] or linear scaling using the largest and smallest peak intensities [33,63,39,41]. Normalization with respect to one or two peaks within a spectrum may be more sensitive to noise than normalization with respect to TIC because the effect of noise at those peaks is transferred to all other peaks through normalization while noise will be canceled out by the summation of peak intensities in normalization with respect to TIC.

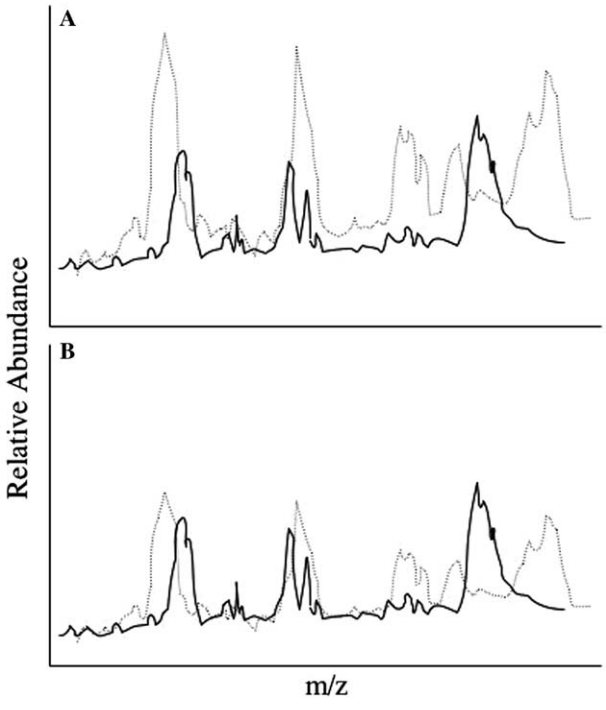

Fig. 2. Normalization is need to compare across spectra since mass spectrometry provides a measure of the *relative* abundance of the different proteins in a sample. In the illustration here, the original spectra (A) are normalized such that the maximum peak heights in each spectra are the same (B).

The four normalization methods described above are performed within a spectrum. Normalization across samples has also been investigated. All the peak intensities at the same mass-to-charge ratio across samples can be normalized with respect to the median peak intensity [73,67] or linearly scaled using the largest and smallest peak intensities [35,125]. Some investigators have extended simple linear scaling by taking the peak variability into consideration [115]. These methods ignore the absolute difference in peak intensities at different mass-to-charge ratios and consider only the difference in the expression levels between cancer and normal samples. Therefore, small peaks can be considered to be as important as large peaks in normalization across samples. However, it should be noted that noise embedded in small peaks can also be amplified by such normalization methods and it still remains unanswered whether peaks belonging to different spectra can be manipulated without any precedent normalization within a spectrum. At present, it is not clear if one normalization method is superior to the others since there have not been any studies in which normalization methods were compared on the same data set.

Some studies have investigated the use of the log transform to reduce the variability of mass spectra [77,124,46,55,65,115,66,42]. However, one should be cautious in using the log transform since it may make it difficult to separate the additive noise component from the original signal. Suppose that mass spectra have additive random noise with zero mean. Such noise can easily be reduced by simple averaging; however, such noise cannot be reduced by simple averaging after a log transform because summation in the log space corresponds to multiplication in the original space. In addition to the log transform, the square root transform has also been investigated as a means of reducing the variability [65].

4.2. Feature extraction

Features are variables constructed from preprocessed data to summarize the properties of the data [126,127] and the process of constructing features is called as "feature extraction." In decision support systems utilizing mass spectra, feature extraction can be defined as a process of extracting summary information reflecting the

pathological status of a sample from preprocessed mass spectra.

The simplest approach to feature extraction from mass spectra is to use the abundance (intensity) information of *every* m/z measured as the features [59,37–39,41–43]. While this approach to feature extraction is straightforward, it places additional demand on the feature selection and classification stages since a very large number of features are used (≈15,000) and most studies employ a modest number of cases (<500). Moreover, mass spectrometers can only distinguish the masses of proteins within a finite resolution level. More than one m/z measured can correspond to the same protein. Thus, high levels of correlation are expected between close m/z values.

Some studies have employed binning to extract features from the raw mass spectra [58,33,63,48,51]. The m/z points are grouped into a number of bins and a feature is derived from each bin by calculating the average [63] or the maximum peak intensity [58]. The spacing of bins is usually uneven because the number of peaks is not uniformly distributed [63]. Binning is the simplest form of peak detection and alignment, which will be discussed in depth from the next paragraph, in a sense that bins are defined over the m/z axis but they are initially placed at fixed positions across multiple spectra and never adjusted again. Binning is also fairly straightforward to use; however, care must be taken in determining the size and location of bins because improper binning may lead to producing incorrect features, which do not enough reflect the pathological status of samples.

Since abundance data from within the mass error rate are considered to represent the same protein, features are often extracted from mass spectra based on the properties of "peaks" that are comprised of multiple m/z points. In this approach, feature extraction consists of three main components: peak detection, peak alignment, and calculation of feature metrics. Often, commercial software provided with mass spectrometers (e.g., Ciphergen's SELDI-TOF system) and in-house algorithms are combined in the feature extraction process.

The identification of peaks in a mass spectrum is complicated by the error in measuring the abundance as well as the mass error rate. The goal of peak detection is to identify sets of m/z values which comprise "peaks" that are higher than the noise level of a mass spectrum. In many studies, commercial software has been used to find as many peaks as possible and a predefined threshold has been applied to select peaks far higher than the noise level. For example, Ciphergen ProteinChip software detects peaks based on the signal-to-noise ratio (S/N). The S/N is an indicator of how much a peak is distinguished from background noise. If the S/N of a peak is 10, the peak has an intensity value 10 times larger than background noise. Ciphergen ProteinChip software first selects peaks with a high signal to noise ratio (e.g., S/N $\geqslant$ 10) within individual mass spectra. Then, across mass spectra, it finds more peaks with a moderately high S/N (e.g., S/N $\geqslant$ 2) [114,128]. Some researchers have explored alternative peak detection algorithms for more rigorous peak finding [96,124,62,66,53]. Most peak detection algorithms find local maxima within a certain mass-to-charge ratio range and choose the local maxima higher than a threshold of the noise level as peaks [96,124,69,53]. Local maxima of a mass spectrum are located by finding the mass-to-charge ratios with the highest intensity among their N neighbors [96,53].

Clearly, peak detection algorithms must include a definition of the noise level around a local maximum. The noise level is often defined as the average of the intensities at the mass-to-charge ratios within a moving window with a fixed size (e.g., 5% of all mass-to-charge ratios in a mass spectrum) [53] or as the median elevated level from the median difference of all local maxima and their adjacent local minima in a mass spectrum [96].

Peak detection, as described above, is concerned with identifying peaks *within a single mass spectrum*. However, to make inferences about trends across several spectra, one must relate the peaks identified in one spectrum to the peaks identified in another spectrum. This process of matching peaks that represent the same protein specie *across multiple spectra* is referred to as "peak alignment" (Fig. 3). In peak alignment, the peaks of multiple mass spectra within the mass error rate are grouped together and regarded as a "peak group."

Most peak alignment algorithms group the peaks around a prominent peak within a moving window the size of the mass error rate in a mass spectrum. Then, the peak groups within the mass error rate are regrouped across spectra and the members of a group are adjusted [44,74,77,36,124,46,67,69,52,119]. In one study, a genetic algorithm was employed to optimize the process of window-based peak alignment [70]. Peak alignment simply based on the mass error rate can produce peak groups that cannot effectively represent proteins in a complex sample. A genetic algorithm was used to identify the peaks that were present across the most samples while at the same time avoid those that were within the mass error rate of those already selected.

After peak detection and peak alignment, one must define the metrics of a peak group that will serve as features. Feature metrics related to peak heights have been used in most studies. The maximum peak height [64,39], average peak height [74,63], and median peak height of a peak group [68] have been used. Instead of retaining the peak height as continuous feature data, binary [53] and discretized feature [60] values have also been investigated as a way to alleviate the variability of feature values across samples that can deteriorate the generalization of the classifier. Binary feature values indicate whether a peak is expressed over the noise level

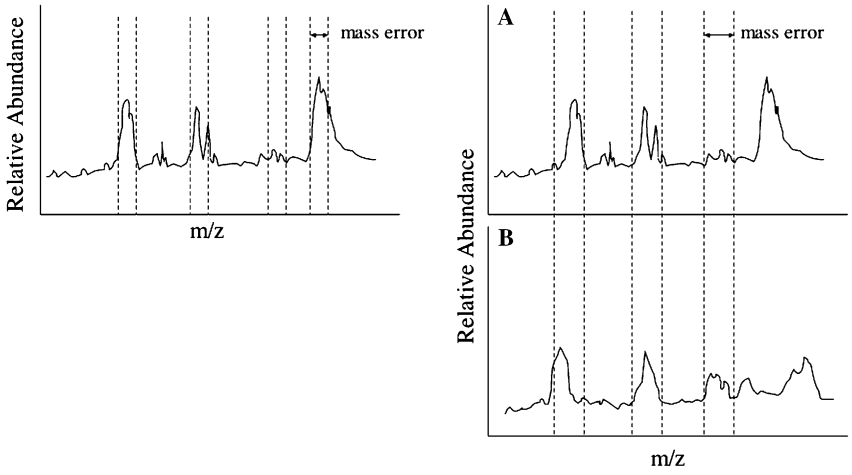

Fig. 3. The left panel illustrates peak detection, which is concerned with identifying peaks *within a single mass spectrum*. The right panel illustrates the process of matching peaks that represent the same protein species *across multiple spectra* (A and B), which is referred to as "peak alignment."

and further discretized feature values specify the degree to which a peak is expressed. Some studies have employed the sum of peaks in a peak group, i.e., the ion current of a peak group, to take into account the contributions of all the peaks representing one protein [35,64].

Most feature extraction methods, as described above, extract features from signals in the original space, i.e., peak intensities of mass spectra. In a few studies, features were extracted by projecting the signals from the original space onto another, usually lower-dimensional, space through linear transformations. Principal component analysis (PCA) has been widely used as a standard way for this purpose in many other data mining applications [129]. PCA identifies the orthogonal directions in which data vary maximally using the eigenvalue/eigenvector decomposition of the covariance matrix. Then the original signals are projected onto those directions, the number of which is usually smaller than the original dimension. The projections are called principal components and often used as features. Since only those directions that explain data variation maximally are selected in PCA, the projected data is of a lower dimension, but with a minimum loss of information. In one study, every m/z point was regarded as a dimension and PCA was applied to find principal components, which were used as features in clustering analysis [61]. The WT has been also employed not only to reduce noise but also to extract features from mass spectra in a similar fashion as PCA is used [50,71]. The WT also compresses data by projecting the original data onto *prespecified* orthogonal directions (wavelets). The coefficient of each wavelet becomes a feature in this case [50,71]. Since the wavelets representing high frequency components are usually ignored, noise reduction is simultaneous accomplished with feature extraction. Both approaches are very sensitive to the choice of components (i.e., principal eigenvectors in PCA or wavelets in the WT); therefore, it is important to determine criteria for selecting eigenvectors or wavelets prior to feature extraction. However, this is currently performed in an ad hoc manner. In addition, as compared with methods that select features in the original space, the features resulting from PCA or the WT are less interpretable because the features are extracted from the *projected* space. Thus, the inverse transformations are needed to reveal how features (m/z points) in the original space contribute to creating each feature in the projected space.

In feature extraction, a variety of peak detection and alignment algorithms are being developed and tested. The resolution and noise of mass spectrometry systems should be taken into account. For example, using the maximum peak of a peak group might lead to over/underestimation of relative abundance of a certain protein because it can be easily affected by noise. Likewise, peak alignment that only considers the mass error rate might deteriorate the sensitivity and specificity. It is possible that better diagnostic systems could be developed if more prior knowledge of mass spectrometry and the proteins present in blood was incorporated into the feature extraction process.

4.3. Feature selection

The purpose of feature extraction is to produce a set of quantitative measures from a mass spectrum that could potentially be used for distinguishing spectra of normal and cancer samples. Typically, the feature extraction process results in a smaller set of features (<1000) than the number of (m/z, relative abundance) pairs that were in the original spectrum ($\approx 15{,}000$). However, in general, the number of features extracted is still much larger than the number of samples (<500) in an experiment. This imbalance in the number of features and samples may increase the chances of misclassifi-

cation due to overtraining and the usage of irrelevant or redundant features [130,131,127,132,40]. Also, a large number of features usually lead to an increase in the training time of classifiers. Moreover, from a biomedical perspective, it is important to find a moderate number of proteins that most contribute to correct classification such that these potential biomarkers can be identified and biochemically validated. Thus, it can be important to reduce the number of features from the set initially extracted. This process is referred to as feature selection.

Feature selection is defined as a series of actions to choose a subset of features that are relevant to correct classification based on specified evaluation and selection criteria [131,127,132,38]. Feature selection methods are often categorized as filters, wrappers, or embedded methods (Fig. 4). A filter method evaluates and ranks individual features based on selection criteria (e.g., t statistic). Then, a subset of features for classification is determined based on individual feature ranks. Wrappers assess the relevancy of a subset of features based on evaluation metrics of a classifier trained using that subset of features. A search algorithm is used to explore the space of feature subsets and identify a high-performing subset of features. Cross-validation or bootstrap sampling are used in conjunction with wrapper methods since they can provide the unbiased accuracy estimates of the classifier. Embedded methods implicitly perform feature selection as a part of the classifier training process.

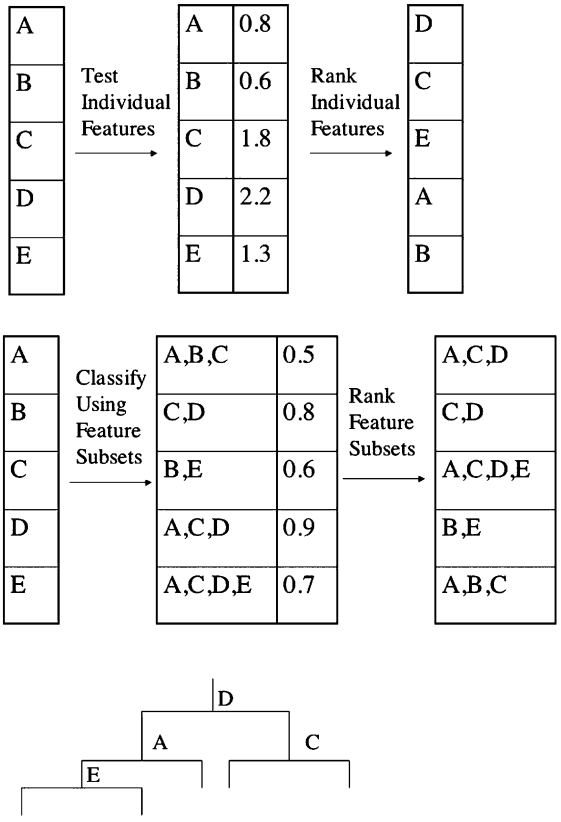

Fig. 4. Feature selection methods are often categorized as filters (top panel), wrappers (middle panel), or embedded methods (bottom panel). A filter method evaluates and ranks individual features based on selection criteria (e.g., t statistic). Then, a subset of features for classification is determined based on individual feature ranks. Wrappers assess the relevancy of a subset of features based on evaluation metrics of a classifier trained using that subset of features. Embedded methods implicitly perform feature selection as a part of the classifier training process (e.g., decision tree).

Filters have been the most commonly used type of feature selection in prior studies of cancer classification using mass spectra. A variety of statistical tests have been investigated to define selection criteria for the relevancy of individual features. The two-sample t test has been used in many studies [59,75,76,36,73,133,43]. A t test for two independent samples (cancer, normal) is performed on each feature across the training samples and features that show a statistically significant difference (e.g., $p < 0.05$) in the group means are selected for use in training classifiers. Other studies have also used methods related to the t test for two independent samples. Li et al. [38] define the distance between two sample groups, cancer and normal, as the absolute mean difference normalized by the root mean square of the variances of two sample groups. This distance measure resembles the two-sample t test for independent samples with unequal variance. Zhu et al. [43] calculated a reliable threshold for p value based on 1D Gaussian random field considering the fact that multiple comparisons are made. Other types of statistical tests such as the χ^2 test [79,133], the one-way analysis of variance (ANOVA) [69,52], the Wilcoxon signed rank test [74,45,41,72], and the Mann–Whitney test [72] have also been used to rank features.

Some studies have tested the efficacy of relevancy measures on the basis of information theory and signal processing as filters. Information gain and relief-F [132,134] are examples of measures used in information theory based filters [60]. The wavelet transform can also be used as a filter method for feature selection. In one study, features were assessed by comparing the wavelet coefficients of each feature between cancer and normal samples [71]. Receiver operating characteristics (ROC) analysis [135] has also been used to measure the relevancy of an individual feature. The area under the curve of each feature is calculated and it is used as the metric to rank features [44]. ROC analysis is discussed further in the evaluation section.

Using a single relevancy measure can lead to biased feature selection. Thus, combinations of methods have been investigated for feature selection [74,36,70]. A feature is considered to be relevant when the feature receives high scores from multiple methods. This approach enables one to explore features from different perspectives and to make a more reliable decision regarding the selected subset of features.

Wrappers are different from filters in that classifier evaluation metrics are used rather than selection criteria

for individual features and wrappers assess features in groups rather than individually. Filters employ selection criteria such as statistical tests to evaluate individual features, while wrappers use evaluation metrics of classifiers to estimate the discriminating power of a candidate subset of features [130,131,127,132]. Moreover, while filters simply select a subset of features by choosing those that were highly ranked individually, wrappers iteratively optimize the subset selection using search algorithms such as genetic algorithms and stepwise selection methods [130,131,127,132]. The wrapper approach typically has better performance than the filter approach since the search process in wrappers enables it to exclude redundant features when forming a subset of features [127,132]. However, the filter approach does have the advantage that is less computationally demanding than the wrapper approach [132,38].

Several studies have investigated the efficacy of wrappers for feature selection in mass spectra. The combination of genetic algorithms [134] with classifiers is a popular use of wrappers in this field [58,38,39]. Several kinds of classifiers have been combined with genetic algorithms, including self-organizing maps [33,34,48,39], support vector machines (SVM) [38], and simple distance based classifiers (e.g., Mahalanobis distance) [58,75,50]. In other studies, stepwise feature selection methods (forward selection and backward elimination) [127] have been used instead [64,41]. A wrapper that incorporates unified maximum separability analysis (UMSA) and bootstrap sampling has identified the best performing subset of features in three studies [77,46,55].

Embedded methods implicitly perform feature selection as a part of the classifier training process [127]. For example, decision trees estimate the contribution of individual features to correct classification in each iteration and grow the tree structure according to the estimation result. Therefore, when the training is over, the final subset of features is produced with the classifier [127,134]. Feature selection using embedded methods for mass spectra will be further discussed in the next section on classification.

Feature selection can help to reduce running time and avoid overtraining if it succeeds in finding a subset of independent and discriminating features. Unfortunately, there is no guarantee that the feature selection process will improve the classification performance. Moreover, features selected as relevant for classification still need to be biologically validated in future studies. Efforts to identify the proteins corresponding to relevant features should follow feature selection and classification studies.

4.4. Classifier training

Machine learning is a branch of artificial intelligence that is concerned with design and application of algorithms that enable computers to learn from experience [134]. We interpret this definition broadly to include techniques that were developed from a statistical, rather than computer science perspective, such as linear discriminant analysis and regression.

There are three general types of machine learning algorithms: unsupervised, reinforcement, and supervised. In unsupervised learning, the computer attempts to identify natural groupings within a dataset based on criteria that define how "similar" items are and what makes a "good" group, but without being provided examples of the feature values of items and associated "correct" class membership. For this reason, unsupervised learning methods are also referred to as "clustering." Unsupervised learning algorithms have not been used in many prior studies of cancer diagnosis from mass spectra. Some studies have explored self-organizing maps [33,34,48,39,51] and hierarchical clustering algorithms [124,73,65] in this field. In reinforcement learning, the computer is not provided with examples of the feature values of items and associated "correct" class membership, but is provided less specific feedback that indicates if the system is on the right track. We are unaware of any studies of mass spectra for cancer diagnosis that employ reinforcement learning methods. In supervised learning, the computer is provided with examples of the feature values of items and associated "correct" class membership. The goal of supervised learning is to develop a "classifier" that can predict the class membership from a set of pre-determined classes for an item based on a set of features that describe the item. Supervised learning methods have been used extensively in the investigation of cancer diagnosis from mass spectra. Prior studies have tested the performance of several supervised learning algorithms including artificial neural networks (ANN) [82,60,73,125,68,136], k nearest neighbor (KNN) [60,125,69,52,43], logistic regression [77,36,46,55,64], decision trees [44,54,60,63,64,125,80,71], linear or quadratic discriminant analysis (LDA/QDA) [58,47,64,50,66,69,52,42], support vector machines (SVMs) [38,125,69,42], kernel matching pursuit (KMP) [62], logical analysis of data (LAD) [32], stepwise discriminant analysis [41], partial least square projection [61,65], Naïve Bayes [60], rule induction [60], and ensemble algorithms (e.g., boosting, bagging, or random forest) combined with various base classifiers [37,49,42,53]. Two evolving themes in the use of supervised learning in this field are the emphases on SVMs and ensemble methods.

SVM is a fairly new class of supervised machine learning methods that has generated considerable excitement (Fig. 5). SVMs are a type of kernel learning methods, which project data from the current vector space to another vector space where linear learning algorithms can be applicable. The functions that project the data onto the new vector space, which usually has a higher

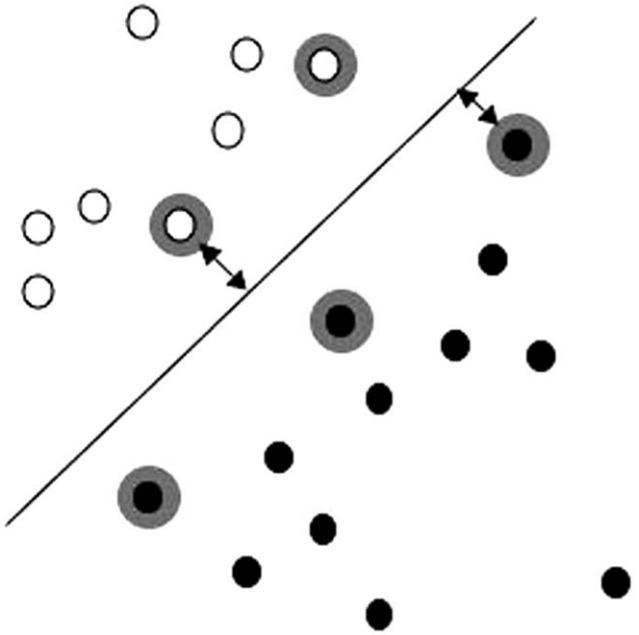

Fig. 5. SVMs are a type of kernel learning methods, which project data from the current vector space to another vector space where linear learning algorithms can be applicable. SVMs guarantee the maximal margin between cancer and normal samples through global optimization of the decision boundary such that overtraining can easily be avoided (margin indicated by arrow). Since the decision boundary set by SVMs has the gradient that allows for the maximum-margin separation based on a few data samples closest to the decision boundary, which are called support vectors (highlighted with gray), SVMs implicitly reflect the contribution of each feature to successful classification and reduce the effect of irrelevant features by performing the dot product between the gradient and each sample.

dimension than the original, are called the kernel functions. Since an improper kernel function may worsen classification by projecting data onto a space where linear separation is impossible, care must be taken for choosing a kernel function when using SVMs. Unfortunately, there are no guidelines for choosing the best kernel for a given data set. Prior knowledge of the characteristics of data may help with this process, but in practice selecting an optimal kernel remains a significant challenge. The most popularly used kernel functions are the polynomial, radial basis function, and sigmoid kernels.

After data projection into a linear space, SVMs guarantee the maximal margin between cancer and normal samples through global optimization of the decision boundary such that overtraining can easily be avoided [137,138]. In the cases where the projected data is still not linearly separable, for example, when two classes overlap, a penalty is given to the objective function of optimization to trade the margin size and misclassification rate. A small penalty maximizes the margin size but increases the misclassification rate while a large one decreases the margin size but minimizes the misclassification rate [139].

SVMs can also be utilized without any data projection if the data are linearly separable in the current vector space. This method is usually called linear-SVMs. Since the decision boundary set by SVMs has the gradient that allows for the maximum–margin separation based on a few data samples closest to the decision boundary, which are called support vectors, SVMs implicitly reflect the contribution of each feature to successful classification and reduce the effect of irrelevant features by performing the dot product between the gradient and each sample. There is less need for an effective feature selection step when a classifier that is robust to irrelevant features is used. The robustness of SVMs to irrelevant and redundant features is especially valuable since mass spectra data sets typically have many more features than cases. Thus, SVMs exhibit several properties that are appealing in the analysis of mass spectra.

The complexity and subtlety of mass spectra patterns between cancer and normal samples may increase the chances of misclassification when a single classifier is used because a single classifier tends to cover patterns originating from only part of the sample space. Therefore, it would be beneficial if multiple classifiers could be trained in such a way that each of the classifiers covers a different part of the sample space and their classification results were integrated to produce the final classification.

Ensemble algorithms such as bagging, boosting, or random forests improve the classification performance by associating multiple base classifiers to work as a "committee" for decision-making [140,141]. Any supervised learning algorithm can be used as a base classifier. Ensemble algorithms not only increase the classification accuracy, but also reduce the chances of overtraining since the committee avoids a biased decision by integrating the different predictions from the individual base classifiers.

Feature selection has been performed as an "embedded" part of the training process in many studies, especially when decision tree or SVM methods were used. Decision trees select the most discriminant features based on the information gain at each stage when growing the tree structure. As a result, a list of features that make the largest contributions to successful classification are obtained when classifier training is finished. In some studies of cancer classification using mass spectra, features selected implicitly by decision trees have been proposed as potential biomarkers [44,49,80]. SVMs also possess embedded feature selection mechanisms. As described in the earlier part of this section, the decision boundary includes the information of each feature's relevancy for successful classification. For example, in the case of the linear SVM, the absolute magnitude of coefficients of the decision boundary (a hyperplane) corresponds to the degree of relevancy of features. Prados et al. [125] proposed a list of potential biomarkers using the internal feature selection function of a linear SVM.

The goal is to build *reliable* classifiers, which can classify unknown samples within a reasonably bounded error range. While the error of a classifier on the training set decreases as the training process proceeds, the error on the general population increases after a certain time point in the training process because the classifier becomes oversensitive to the patterns that exist only in the training set. This event is called as "overtraining." It is important to avoid overtraining by evaluating the classifier performance using an independent set of samples. In addition, it is impossible to find a classification algorithm superior to the others for all feature selection methods because every classification algorithm has its own learning bias [126,134]. The performance of a classification algorithm can be varied by the choice of feature selection methods. For example, KNN is very sensitive to irrelevant and redundant features. However, in prior studies, the relationship between the chosen feature selection method and classification algorithm has not been thoroughly researched. It is necessary to identify the best pair of a feature selection method and classifier.

4.5. Evaluation

After a system is developed through the stages described in the previous sections, its performance must be carefully assessed. In this section we discuss two important issues in system evaluation. First, the quality of the data set used to develop the system will strongly influence its performance since systems for cancer diagnosis from mass spectra are inherently data-driven. Second, the system evaluation must be based on criteria that are clinically relevant and quantitative with clearly defined standards of interpretation.

The desired characteristics of the data are that they provide an accurate representation of the population to be tested and that there are sufficient data to allow for robust inference. There are many factors that can bias a sample such that it does not correctly describe the population, e.g., the choice of human subject inclusion/exclusion criteria, data entry errors, etc. This problem is complicated by the fact that disease cases typically must be present in the data set at a much higher proportion than the population prevalence in order to show the breadth of variability in the disease state with a limited overall sample size. The imbalance in the sizes of disease and healthy classes can make classifiers more sensitive to patterns originating from disease cases, resulting in more false positives in classification. If there are more healthy cases, the number of false negatives will increase because patterns from healthy cases will be relatively more emphasized. From this point of view, it is valuable to equalize the class sizes [142–144]. However, there would be difficulty in keeping the balance between disease and healthy sample sets as one attempts to increase the entire sample size for more robust and reliable inference because disease cases are usually more difficult to obtain than healthy ones. To the best of our knowledge, this issue has not yet been addressed in the arena of analyzing mass spectra. Over-sampling the minority class and under-sampling the majority class have been common methods to resolve biased classification due to imbalanced data. The basic idea behind these techniques is to balance the sizes of two classes artificially. For example, over-sampling the minority class, i.e., sampling with replacement, increases the size of the minority class up to that of the majority class. Similarly, under-sampling the majority class, i.e., decimating samples, can reduce the size of the majority class up to that of the minority class. However, it should be noted that these two techniques must be carefully used because over-sampling a minority class may lead to overtraining to a specific pattern of the samples belonging to the minority class and under-sampling a majority class may lose some valuable patterns of majority class samples [142,144]. Some studies have tried to resolve this issue by penalizing the error rates of the samples of the minority class more, which prevents the classifier from sacrificing those samples of the minority class to decrease the overall error rates (e.g., 1-accuracy) [142–144].

Proper handling of mislabeled data samples is also an important issue for classifier training and evaluation. There are two approaches to contending with mislabeled data. One is to reduce the likelihood of its existence through experimental design and quality control. The second is to eliminate mislabeled data in post hoc fashion during the analysis. In practice, since even extremely rigorous experimental design and quality control may not be able to perfectly prevent the occurrence of mislabeled data, both approaches should be taken to alleviate the effects of mislabeled data on decision support systems [145–147].

To avoid mislabeled data through experimental design and quality control, we must consider the possible sources. For example, mislabeling can arise from data entry errors. To a large extent, this can be avoided through rigorous laboratory protocols. A more concerning source of mislabeled data is genuine confusion regarding the correct classification of a sample due to the error or limitations inherent to the diagnostic test used to establish truth or the absence of a test for truth. For example, a healthy sample may be mislabeled as positive based on a false-positive biopsy. This type of error can be avoided if samples are only included for study if they have undergone confirmatory testing (e.g., repeated biopsy). On the other hand, a diseased sample can be mislabeled as healthy either because of a false-negative diagnostic test or because no diagnostic testing was performed (e.g., an asymptomatic subject was presumed to be healthy). Given the limitations of existing diagnostic tests for detecting very early stage disease and the many

reasons not to perform diagnostic tests on seemingly healthy individuals, this can be an important source of false-negative samples. The most common approach to avoid this problem is to only consider a healthy sample to be healthy after an appropriate duration of disease-free follow-up time [148]. To the best of our knowledge, this issue has not been explicitly discussed in any reports of studies of cancer classification from mass spectrometry to date. Moreover, we are unaware of any studies that have demonstrated and analyzed the risk of system performance degradation due to mislabeled training/test data samples in the context of mass spectrometry analysis.

The machine learning literature can provide some guidance on post hoc methods for detecting mislabeled samples. Mislabeled samples may appear as outliers. Therefore, detecting mislabeled samples is closely related to detecting outliers. Some approaches for outlier detection have been developed. For example, simply analyzing the means and standard deviations of features with the confidence intervals of each feature can reveal outliers [145] because samples lying outside the confidence interval are highly probable to be outliers. Clustering algorithms also can be used to identify outliers [149,145]. Presumably, samples belonging to the same class would be clustered together while outliers would behave as ones belonging to other classes. Note that this clustering should be performed prior to feature selection. Other studies have used multiple classifiers of different types to filter out outliers [150,151]. The key idea is that the samples whose labels were consistent with the labels predicted by multiple classifiers were regarded as correct samples and that were not were regarded as outliers.

There is no theory to provide firm guidance on the sample sizes required to properly perform any of the stages of development of clinical decision support systems utilizing mass spectrometry of blood products. Sometimes, it is easy to identify in retrospect that a sample may have been too small, such as when an algorithm fails to converge or operates with unacceptably low performance. However, one needs to take care in devising evaluation strategies that help avoid the common and difficult problem of the system appearing to perform well on the data set used for development but proving unsatisfactory when subjected to additional testing with more data. Fortunately, this danger can be reduced to a large degree through appropriate use of data partitioning and sampling schemes.

In general, three independent sets of samples are needed for the development and evaluation of a classification system [134]. One set is called the training set and used for training a classifier. During or after classifier training, the classifier should be pruned and adjusted to avoid possible overtraining using another, independent sample set, which is referred as the validation set. As described in the previous section on classifier training, the error on the validation set tends to increase after a certain time point while the error on the training set keeps decreasing as the training process continues. The time point at which the error on the validation set starts to increase is the point when training should conclude. The validation set is used to find the stopping point of training. After the classifier is developed using the training and validation sets, it must be evaluated with respect to the general population. The test set is used to estimate the true error of the classifier on the general population. It is also important to recognize that a mass spectrometry analysis is actually composed of a series of chemical/biochemical processes. Thus, within a data set samples must be randomized in each analytical step so as to avoid any possible bias due to batch processing because such bias could produce systematic patterns that interfere the ''true'' patterns originating from the pathological changes in the samples.

Typically, the same data (training set) are used in the procedures of preprocessing, feature extraction, feature selection, and classifier training. The use of separate sets for choosing algorithms and setting their parameters in each of these stages would provide greater protection against overtraining. Unfortunately, this is seldom plausible given realistic sample sizes. In fact, in most studies of cancer classification using mass spectra, the number of available samples is not even large enough to produce three independent sample sets. Even when three non-overlapping sets are used, they are typically partitioned from a single set and as such as are not truly ''independent'' sets. We are aware of very few studies of cancer classification using mass spectra of human blood samples that have employed a truly independent test set (e.g., test set was generated on a different day than the training set) [43].

The small number of cases necessitates the use of sampling techniques such as k-fold cross-validation, bootstrap sampling [152–154], or random partitioning (Fig. 6) to estimate the generalization ability of the classifier. Sampling techniques are used to obtain estimates of classifier performance by judicious reuse of data. However, it should be noted that no sampling technique can perfectly address the question of how systematic and realistic variations in the data source (e.g., variations in a single mass spectrometer over time or between two mass spectrometers) will impact the general classifier performance. A classification system must ultimately be evaluated using a large, independent data set.

In k-fold cross-validation, the data are split into k non-overlapping subsets or ''folds'' such that each sample is present in a single fold [152]. The classifier is trained on $k-1$ of the folds and tested on the remaining fold. This process is repeated such that each fold is withheld once. Usually, the average of the evaluation results (e.g., accuracies) across the folds is taken as the estimate

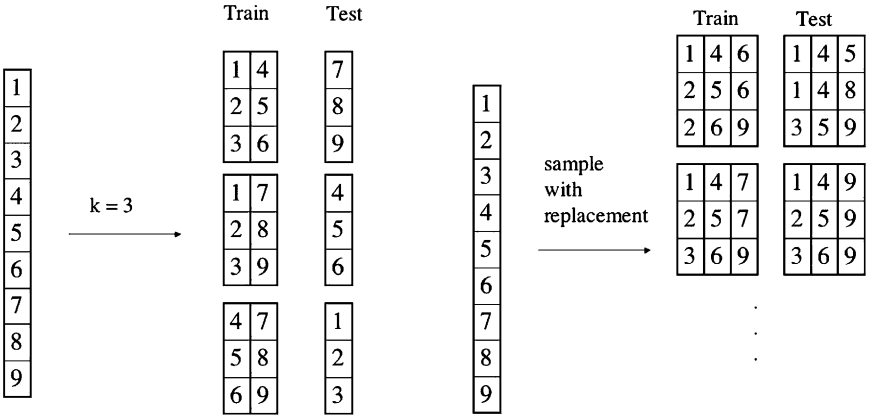

Fig. 6. The left panel illustrates k-fold cross-validation and the right panel illustrates bootstrap sampling. In k-fold cross-validation, the data are split into k non-overlapping subsets or "folds" such that each sample is present in a single fold. The classifier is trained on $k-1$ of the folds and tested on the remaining fold. This process is repeated such that each fold is withheld once. Usually, the average of evaluation results (e.g., accuracies) across the folds is taken as the estimate of the overall system performance. By comparison, a bootstrap set is created by random sampling of N cases with replacement from the original set of N cases. A classifier is trained on one such bootstrap set and tested on another. The process is repeated many times and the average of the evaluation results across the bootstrap sets is taken as the estimate of the system performance.

of the overall system performance [32,54,60,34,77,37,55,62,64,73,125,49,50,66,56,52,80,42,71]. When k is equal to the number of samples, this procedure is called leave-one-out cross-validation. In leave-one-out cross-validation, every sample is tested exactly one time and the overall system performance is estimated by simply gathering the individual sample validation results as if these test results came from a single classifier [32,58,61,38,63,65,66,40,70,71,43]. Note that the actual number of classifiers trained is equal to the value of k in the cross-validation.

It is important to remember that the performance estimates obtained by k-fold cross-validation are affected by the size of the training set and the number of folds. An estimator is evaluated in terms of its bias, the extent to which the average system performance estimate is close to the true system performance in the population, and its variance, the extent to which the estimates spread around the average system performance estimate [155]. The estimate of the true system performance is more biased as the size of the training set decreases and has higher variance as the size of the testing set decreases [156,157]. Therefore, a cross-validation using a larger value for k will result in an estimate with less bias, but higher variance relative to a cross-validation using a smaller value for k. Several excellent texts are available that discuss the trade-offs between bias and variance in classifier evaluation [155,126,156,157].

The bootstrap sampling is another technique to estimate the true system performance with a limited number of samples [152,153]. A bootstrap set is created by random sampling of N cases with replacement from the original set of N cases. A classifier is trained on one such bootstrap set and tested on another. The process is repeated many times and the average of the evaluation results across the bootstrap sets is taken as the estimate of the system performance [34,42]. Note that each bootstrap set created for training results in a separate classifier. One study [42] employed 0.632+ bootstrap sampling, a modified version of bootstrap sampling, which can alleviate the bias in estimating the true system performance [157].

Random partitioning can be regarded as single or multiple 2-fold cross-validation. It is also similar to bootstrap sampling except that sampling is performed without replacement and less than N of N cases are selected. The training set is generated by randomly sampling a certain portion of data and the remaining samples of data are used as the test set [44,74,34,36,47,64,48,39,49,50,41,51,56,70,53].

Commonly, cross-validation, bootstrap sampling, and random partitioning are used to estimate the system performance during the classifier training stage. However, some studies have applied random partitioning to derive reliably discriminant features during feature selection [74,77,46,55] based on the ranks of discriminant features that are earned on each sampled training data set. The features with consistently high ranks are selected for use. However, in practice, it is often impossible or very difficult for the *entire* design process to be performed in a cross-validation manner. As a consequence, several previous studies seem to have used the full data set prior to cross-validation for feature selection [77,47,63,73,66,69,42]. As was the case for estimating system performance, sampling techniques cannot overcome limitations that are inherent to the data set from which the samples are drawn. If the data set does not represent the underlying probability distribution of the population of interest, then even the most sophisticated feature selection based on sampling techniques

will end up with an extremely "biased" subset of features [148].

It is critical that the system evaluation be based on criteria that are clinically relevant and be quantitative with clearly defined standards of interpretation. While classifiers typically attempt to optimize an evaluation function as part of the training process, it is important recognize that in general that function is not the most clinically relevant measure. For example, the mean-square error measure weights the two possible kinds of error equally while in most medical diagnostic tasks the costs, monetary and otherwise, of false-positives and false-negatives are not equal.

Accuracy, the fraction of the samples that the system correctly classifies, has been used in many mass spectrometry studies that employ a binary decision approach [58,82,60,36,136,61,37,38,47,62–64,125,65,81,40,69,52,42,70,71]. However, there is a significant drawback to the accuracy metric in that it is dependent on the prevalence of disease in the data set. For example, if there are only 20 disease cases for every 80 normal cases, a system could achieve 80% accuracy by simply reporting all cases as normal. Thus, if the prevalence is not 50%, the system accuracy cannot be interpreted in isolation. The most clinically relevant measures for screening and diagnostic tests are sensitivity and specificity, regardless of whether the test involves a computational aid. Many studies of mass spectrometry for cancer classification have used these measures [44,32,54,74,45,33,97,75,77,36,46,55,38,47,63,39,73,125, 49,50,66,41,51,56,133,69,80,53,43].

Receiver operating characteristic (ROC) analysis can be used for diagnostic systems that provide a range of outputs rather than a binary classification. An ROC curve is a plot of the sensitivity vs. (1-specificity), or equivalently the true positive fraction vs. the false positive fraction, computed from the application of a series of thresholds to the system output (Fig. 7). The advantage of ROC analysis is that it explicitly shows the trade-offs in sensitivity and specificity that could be achieved with the same classification system. In essence, the choice of the decision threshold is delayed until a later time when more knowledge may be available on the costs associated with each type of error.

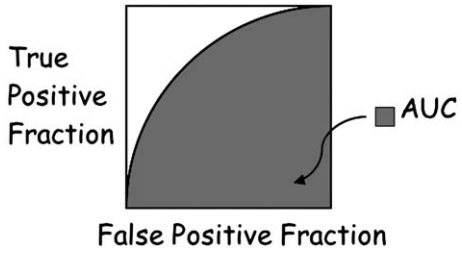

Fig. 7. Receiver operating characteristic (ROC) analysis can be used to evaluate diagnostic systems that provide a range of outputs rather than a binary classification. An ROC curve is a plot of the sensitivity vs. (1-specificity), or equivalently the true positive fraction vs. the false positive fraction, computed from the application of a series of thresholds to the system output. A measure of the concaveness of ROC curves is the area under the curve (AUC).

In general, ROC curves are concave and better system performance corresponds to more concave curves. A measure of the concaveness of ROC curves is the area under the curve (AUC). Hence, the AUC has been used as a measure of system performance in many studies [74,77,36,136,55,38,63,125,66,80].

Evaluation metrics (e.g., ROC AUC) are calculated based on a given data samples, yet it is the performance on the general population that matters. Therefore, there is a need to estimate the reliability of the system. For this purpose, some studies have randomly permuted the class labels of samples and compared the performance to that from using the actual class labels [61,64,69,52]. As the difference between two becomes larger, the performance evaluation from the actual samples is taken as a more reliable indicator of how the system would perform on the general population.

When sampling techniques are used, care must be taken not to mistakenly tune classifier performance results on the "testing" portions. For example, several studies appear to have determined the threshold for calculating the sensitivity and specificity based on the "testing" portion of the data rather than the "training" portion [74,77,36,55,38,47,63,73,66]. These practices can partially undermine the protection against overtraining provided by those sample techniques.

The use of appropriate data sampling methods and relevant evaluation metrics can provide substantial reassurance that laboratory studies will contribute towards the goal of accurate and reliable clinical decision support systems. Of course, laboratory studies must be followed by rigorous clinical testing. For example, studies of the way that the healthcare team does, or does not, incorporate the recommendations made by a system based on the mass spectrometry data is beyond the scope of this review. Ultimately, long-term, large clinical trials are required to establish the efficacy of any screening test to the level of a decrease in cause-specific mortality.

5. Summary

An ideal screening method should be accurate, reliable, rapid, inexpensive, and minimally invasive. Proteomic profiling of blood samples using mass spectrometry has recently been proposed as a method that has the potential to meet these goals. However, there are key difficulties that must be addressed before clinical diagnostic tools can be developed based on this technology. Chief among these is to overcome the restrictions on reliability that have plagued early studies. To achieve accurate

classification on a given set of samples is useless unless the classifier can also be generalized such that new, but similar, data can be accurately classified. A system for discriminating proteomic patterns of samples from healthy and ill people must be robust to the variability that will exist across people, mass spectrometers, sample collection protocols, days, etc.

This article reviews the literature on developing clinical decision support systems for cancer screening from proteomic patterns obtained by mass spectrometry of blood samples from a machine learning perspective. Prior studies are presented in an explicit machine learning framework consisting of five stages: preprocessing, feature extraction, feature selection, classifier training, and evaluation. The purpose of preprocessing is to reduce the influence of aspects of the data that are not expected to aid in the goal of discrimination between disease and healthy patterns and instead may make that classification task more difficult. In feature extraction, the aim is to reduce the dimensionality of the data and increase the interpretability by defining numerical summary measures, often called "features." Following feature extraction, it is necessary to perform a feature selection step in which a subset of features that best enable discrimination between the two groups is identified. Given a set of spectra summarized by informative features and with corresponding truth (health status), a variety of classification algorithms can be trained. Finally, care must be taken in the choice of experimental design (e.g., data sampling) and evaluation criteria to assess both accuracy and reliability (generalization).

It is apparent that the components of the framework that are most specific to the data type, mass spectra of blood samples, are preprocessing, feature extraction, and feature selection. We hypothesize that improvements in these components will yield the greatest increase in system reliability and that the approaches most likely to achieve those improvements will be based on explicit models of the data generation. While the objective of developing a clinical decision support system for cancer screening from proteomic patterns is ultimately data driven, we argue that this goal may not be achievable with reasonable sample sizes unless we use knowledge of the related biology, chemistry, and engineering to constrain the design process.

Acknowledgments

The authors thank Kelly N. Forsythe and Nick Markey for bibliographic data entry. The authors also thank Dr. John M. Koomen in the Department of Molecular Pathology at the University of Texas M.D. Anderson Cancer Center for his sincere advice and careful revision on the section of this article entitled "Mass Spectrometry." Of course, any errors are ours alone. This work was supported in part by a seed grant from The University of Texas Center for Biomedical Engineering.

References

[1] Cancer facts and figures 2004. Atlanta: American Cancer Society; 2004.
[2] Jemal A, Tiwari RC, Murray T, Ghafoor A, Samuels A, Ward E, et al. Cancer statistics. CA Cancer J Clin 2004;54:8–29.
[3] Etzioni R, Urban N, Ramsey S, McIntosh M, Schwartz S, Reid B, et al. The case for early detection [review] [86 refs]. Nat Rev 2003;3. Cancer.
[4] Fahey MT, Irwig L, Macaskill P. Meta-analysis of pap test accuracy [see comment]. Am J Epidemiol 1995;141:680–9.
[5] Green BB, Taplin SH. Breast cancer screening controversies. J Am Board Fam Pract 2003;16:233–41.
[6] Lee CH. Screening mammography: Proven benefit, continued controversy. Radiol Clin North Am 2002;40:395–407.
[7] Knutzen AM, Gisvold JJ. Likelihood of malignant disease for various categories of mammographically detected, nonpalpable breast lesions. Mayo Clin Proc 1993;68:454–60.
[8] Kopans DB. The positive predictive value of mammography. Am J Roentgenol 1992;158:521–6.
[9] Humphrey LL, Helfand M, Chan BKS, Woolf SH. Breast cancer screening: a summary of the evidence for the US. Preventive services task force. Ann Intern Med 2002;137:347–60.
[10] Walsh JME, Terdiman JP. Colorectal cancer screening: clinical applications. J Am Med Assoc 2003;289:1297–302.
[11] Walsh JME, Terdiman JP. Colorectal cancer screening: scientific review. J Am Med Assoc 2003;289:1288–96.
[12] Pignone M, Rich M, Teutsch SM, Berg AO, Lohr KN. Screening for colorectal cancer in adults at average risk: a summary of the evidence for the US. Preventive services task force. Ann Intern Med 2002;137:132–41.
[13] Rennert G, Rennert HS, Miron E, Peterburg Y. Population colorectal cancer screening with fecal occult blood test. Cancer Epidemiol Biomarkers Prev 2001;10:1165–8.
[14] Vernon SW. Participation in colorectal cancer screening: a review [see comment]. J Natl Cancer Inst 1997;89:1406–22.
[15] Peek ME, Han JH. Disparities in screening mammography. Current status, interventions and implications. J Gen Intern Med 2004;19:184–94.
[16] Brawer MK. Prostate-specific antigen: Current status. CA Cancer J Clin 1999;49:264–81.
[17] Liotta LA, Ferrari M, Petricoin E. Written in blood. Nature 2003;425:905.
[18] Pusch W, Flocco MT, Leung SM, Thiele H, Kostrzewa M. Mass spectrometry-based clinical proteomics [review] [68 refs]. Pharmacogenomics 2003;4:463–76.
[19] Srinivas PR, Srivastava S, Hanash S, Wright Jr GL. Proteomics in early detection of cancer. Clin Chem 2001;47:1901–11.
[20] Wulfkuhle JD, Liotta LA, Petricoin EF. Proteomic applications for the early detection of cancer. Nat Rev Cancer 2003;3:267–75.
[21] Woolas R, Xu F, Jacobs I, Yu Y, Daly L, Berchuck A, et al. Elevation of multiple serum markers in patients with stage i ovarian cancer. J Natl Cancer Inst 1993;85:1748–51.
[22] Abbott A. A post-genomic challenge: learning to read patterns of protein synthesis. Nature 1999;402:715–20.
[23] Madi A, Pusztahelyi T, Punyiczki M, Fesus L. The biology of the post-genomic era: the proteomics. Acta Biol Hung 2003;54:1–14.
[24] Verma M, Wright Jr GL, Hanash SM, Gopal-Srivastava R, Srivastava S. Proteomic approaches within the nci early detection research network for the discovery and identification of cancer biomarkers. Ann NY Acad Sci 2001;945:103–15.

[25] Rai AJ, CHan DW. Cancer proteomics: serum diagnostics for tumor marker discovery. Ann NY Acad Sci 2004;1022:286–94.
[26] Rodland KD. Proteomics and cancer diagnosis: the potential of mass spectrometry. Clin Biochem 2004;37:579–83.
[27] Conrads TP, Zhou M, Petricoin 3rd EF, Liotta L, Veenstra TD. Cancer diagnosis using proteomic patterns. Expert Rev Mol Diagn 2003;3:411–20.
[28] Krieg RC, Paweletz CP, Liotta LA, Petricoin 3rd EF. Clinical proteomics for cancer biomarker discovery and therapeutic targeting. Technol Cancer Res Treat 2002;1:263–72.
[29] Petricoin EE, Paweletz CP, Liotta LA. Clinical applications of proteomics: proteomic pattern diagnostics. J Mammary Gland Biol Neoplasia 2002;7:433–40.
[30] Petricoin EF, Fishman DA, Conrads TP, Veenstra TD, Liotta LA. Lessons from kitty hawk: from feasibility to routine clinical use for the field of proteomic pattern diagnostics. Proteomics 2004;4:2357–60.
[31] Rosenblatt KP, Bryant-Greenwood P, Killian JK, Mehta A, Geho D, Espina V, et al. Serum proteomics in cancer diagnosis and management. Ann Rev Med 2004;55:97–112.
[32] Alexe G, Alexe S, Liotta LA, Petricoin EF, Reiss M, Hammer PL. Ovarian cancer detection by logical analysis of proteomic data. Proteomics 2004;4:766–83.
[33] Conrads TP, Fusaro VA, Ross S, Johann D, Rajapakse V, Hitt BA, et al. High-resolution serum proteomic features for ovarian cancer detection. Endocr Relat Cancer 2004;11:163–78.
[34] Jeffries N. Performance of a genetic algorithm for mass spectrometry proteomics. BMC Bioinformatics 2004;5:180.
[35] Johann Jr DJ, McGuigan MD, Tomov S, Fusaro VA, Ross S, Conrads TP, et al. Novel approaches to visualization and data mining reveals diagnostic information in the low amplitude region of serum mass spectra from ovarian cancer patients. Dis Markers 2003–2004;19:197–207.
[36] Kozak KR, Amneus MW, Pusey SM, Su F, Luong MN, Luong SA, et al. Identification of biomarkers for ovarian cancer using strong anion-exchange proteinchips: potential use in diagnosis and prognosis. Proc Natl Acad Sci USA 2003;100:12343–8.
[37] Li J, Liu H, Ng S, Wong L. Discovery of significant rules for classifying cancer diagnosis data. Bioinformatics 2003;19:93–102.
[38] Li L, Tang H, Wu Z, Gong J, Gruidl M, Zou J, et al. Data mining techniques for cancer detection using serum proteomic profiling. Artif Intell Med 2004;32:71–83.
[39] Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, et al. Use of proteomic patterns in serum to identify ovarian cancer. Lancet 2002;359:572–7.
[40] Somorjai RR, Dolenko B, Baumgartner R. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. Bioinformatics 2003;19:1484–91.
[41] Sorace JM, Zhan M. A data review and re-assessment of ovarian cancer serum proteomic profiling. BMC Bioinformatics 2003;4.
[42] Wu B, Abbott T, Fishman D, McMurray W, Mor G, Stone K, et al. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. Bioinformatics 2003;19:1636–43.
[43] Zhu W, Wang X, Ma Y, Rao M, Glimm J, Kovach JS. Detection of cancer-specific markers amid massive mass spectral data. Proc Natl Acad Sci USA 2003;100:14666–71.
[44] Adam BL, 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.
[45] Cazares LH, Adam BL, 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;8:2541–52.
[46] Li J, White N, Zhang Z, Rosenzweig J, Mangold LA, Partin AW, et al. Detection of prostate cancer using serum proteomics pattern in a histologically confirmed population [article]. J Urol 2004;171(5):1782–7.
[47] Lilien RH, Farid H, Donald BR. Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. J Comput Biol 2003;10:925–46.
[48] Ornstein DK, Rayford W, Fusaro VA, Conrads TP, Ross SJ, Hitt BA, et al. Serum proteomic profiling can discriminate prostate cancer from benign prostates in men with total prostate specific antigen levels between 2.5 and 15.0 ng/ml. J Urol 2004;172:1302–5.
[49] Qu Y, Adam B-L, Yasui Y, Ward MD, 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;48:1835–43.
[50] Qu Y, Adam BL, Thornquist M, Potter JD, Thompson ML, Yasui Y, et al. Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensionality data. Biometrics 2003;59:143–51.
[51] Stone JH, Rajapakse VN, Hoffman GS, Specks U, Merkel PA, Spiera RF, et al. A serum proteomic approach to gauging the state of remission in wegener's granulomatosis. Arthritis Rheum 2005;52:902–10.
[52] Wagner M, Naik D, Pothen A, Kasukurti S, Devineni R, Adam B-L, et al. Computational protein biomarker prediction: a case study for prostate cancer. BMC Bioinformatics 2004;5:26.
[53] Yasui Y, Pepe M, Thompson ML, Adam BL, Wright Jr GL, Qu Y, et al. A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. Biostatistics 2003;4:449–63.
[54] Becker S, Cazares L, Watson P, Lynch H, Semmes O, Drake R, et al. Surfaced-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;11:907–14.
[55] 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;48:1296–304.
[56] Vlahou A, Giannopoulos A, Gregory BW, Manousakas T, Kondylis FI, Wilson LL, et al. Protein profiling in urine for the diagnosis of bladder cancer. Clin Chem 2004;50:1438–41.
[57] Vlahou A, Schellhammer PF, Wright Jr GL. Application of a novel protein chip mass spectrometry technology for the identification of bladder cancer-associated biomarkers. Adv Exp Med Biol 2003;539A:47–60.
[58] Baggerly KA, Morris JS, Wang J, Gold D, Xiao L, Coombes KR. A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. Proteomics 2003;3:1667–72.
[59] Campa MJ, Wang MZ, Howard B, Fitzgerald MC, Patz Jr EF. Protein expression profiling identifies macrophage migration inhibitory factor and cyclophilin a as potential molecular targets in non-small cell lung cancer. Cancer Res 2003;63:1652–6.
[60] Hilario M, Kalousis A, Muller M, Pellegrini C. Machine learning approaches to lung cancer prediction from mass spectra. Proteomics 2003;3:1716–9.
[61] Lee KR, Lin X, Park DC, Eslava S. Megavariate data analysis of mass spectormetric proteomics data using latent variable projection method. Proteomics 2003;3:1680–6.
[62] Liu Q, Krishnapuram B, Pratapa P, Liao X, Hartemink A, Carin L. Identification of differentially expressed proteins using

MALDI-TOF mass spectra. In: ASILOMAR conference: biological aspects of signal processing; 2003.
[63] Markey MK, Tourassi GD, Floyd CEJ. Decision tree classification of proteins identified by mass spectrometry of blood serum samples from people with and without lung cancer. Proteomics 2003;3:1678–9.
[64] Neville P, Tan P, Mann G, Wolfinger R. Generalizable mass spectrometry mining used to identify disease state biomarkers from blood serum. Proteomics 2003;3:1710–5.
[65] Purohit PV, Rocke DM. Discriminant models for high-throughput proteomics mass spectrometer data. Proteomics 2003;3:1699–703.
[66] Sidransky D, Irizarry R, Califano JA, Li X, Ren H, Benoit N, et al. Serum protein MALDI profiling to distinguish upper aerodigestive tract cancer patients from control subjects. J Natl Cancer Inst 2003;95:1711–7.
[67] Slotta DJ, Heath LS, Ramakrishnan N, Helm R, Potts M. Clustering mass spectrometry data using order statistics. Proteomics 2003;3:1687–91.
[68] Tatay JW, Feng X, Sobczak N, Jiang H, Chen C, Kirova R, et al. Multiple approaches to data-mining of proteomics data based on statistical and pattern classification methods. Proteomics 2003;3:1704–9.
[69] Wagner M, Naik D, Pothen A. Protocols for disease classification from mass spectrometry data. Proteomics 2003;3:1692–8.
[70] Yanagisawa K, Xu BJ, Massion PP, Larsen PH, White BC, Roberts JR, et al. Proteomic patterns of tumour subsets in non-small-cell lung cancer. Lancet 2003;362:433–9.
[71] Zhu H, Yu CY, Zhang H. Tree-based disease classification using protein data. Proteomics 2003;3:1673–7.
[72] Zhukov TA, Johnson RA, Cantor AB, Clark RA, Tockman MS. Discovery of distinct protein profiles specific for lung tumors and pre-malignat lung lesions by SELDI mass spectrometry. Lung Cancer 2003;40:267–79.
[73] Poon TCW, Yip T-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;49:752–60.
[74] Bhattacharyya S, Siegel ER, Petersen GM, Chari ST, Suva LJ, Haun RS. Diagnosis of pancreatic cancer using serum proteomic profiling. Neoplasia 2004;6:674–86.
[75] Koomen JM, Shih LN, Coombes KR, Li D, Xiao L-C, Fidler IJ, et al. Plasma protein profiling for diagnosis of pancreatic cancer reveals the presence of host response proteins. Clin Cancer Res 2005;11:1110–8.
[76] Koomen JM, Zhao H, Li D, Abbruzzese J, Baggerly K, Kobayashi R. Diagnostic protein discovery using proteolytic peptide targeting and identification. Rapid Commun Mass Spectrom 2004;18:2537–48.
[77] Koopmann J, Zhang Z, White N, Rosenzweig J, Fedarko N, Jagannath S, et al. Serum diagnosis of pancreatic adenocarcinoma using surface-enhanced laser desorption and ionization mass spectrometry. Clin Cancer Res 2004;10:860–8.
[78] Valerio A, Basso D, Fogar P, Falconi M, Greco E, Bassi C, et al. MALDI-TOF analysis of portal sera of pancreatic cancer patients: identification of diabetogenic and antidiabetogenic peptides. Clin Chim Acta 2004;343:119–27.
[79] 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 Spectrom 2001;15:2420–5.
[80] Won Y, Song H, Kang TW, Kim J, Han B, Lee S. Pattern analysis of serum proteome distinguishes renal cell carcinoma from other urologic diseases and healthy persons. Proteomics 2003;3:2310–6.
[81] Seraglia R, Ragazzi E, Vogliardi S, Allegri G, Pucciarelli S, Agostini M, et al. Search of plasma markers for colorectal cancer by matrix-assisted laser desorption/ionization mass spectrometry. J Mass Spectrom 2005;40:123–6.
[82] Ball G, Mian S, Holding F, Allibone RO, Lowe J, Ali S, et al. An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. Bioinformatics (Oxford) 2002;18:395–404.
[83] Baggerly KA, Morris JS, Coombes KR. Reproducibility of SELDI-TOF protein patterns in serum: comparing data sets from different experiments. Bioinformatics 2004;20:777–85.
[84] Diamandis EP. Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems. J Natl Cancer Inst 2004;96:353–6.
[85] Diamandis EP. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations [review] [68 refs]. Mol Cell Proteomics 2004;3:367–78.
[86] Diamandis EP. Proteomic patterns in biological fluids: do they represent the future of cancer diagnostics. Clin Chem 2003;49:1272–5.
[87] Diamandis EP, van der Merwe D-E. Plasma protein profiling by mass spectrometry for cancer diagnosis: opportunities and limitations. Clin Cancer Res 2005;11:963–5.
[88] Petricoin EFI, Ornstein DK, Paweletz CP, Ardekani A, Hackett PS, Hitt BA, et al. Serum proteomic patterns for detection of prostate cancer. J Natl Cancer Inst 2002;94:1576–8.
[89] Baggerly KA, Morris JS, Edmonson SR, Coombes KR. Signal in noise: Evaluating reported reproducibility of serum proteomic tests for ovarian cancer. J Natl Cancer Inst 2005;97:307–9.
[90] Mehta AI, Ross S, Lowenthal MS, Fusaro V, Fishman DA, Petricoin 3rd EF, et al. Biomarker amplification by serum carrier protein binding. Dis Markers 2003;19:1–10.
[91] Liotta LA, Lowenthal M, Mehta A, Conrads TP, Veenstra TD, Fishman DA, et al. Importance of communication between producers and consumers of publicly available experimental data. J Natl Cancer Inst 2005;97:310–4.
[92] Grizzle WE, Meleth S, Petricoin EF, Liotta LA. Clarification in the point/counterpoint discussion related to surface-enhanced laser desorption/ionization time-of-flight mass spectrometric identification of patients with adenocarcinomas of the prostate * proteomic pattern complexity reveals a rich and uncharted continent of biomarkers. Clin Chem 2004;50:1475–7.
[93] Semmes OJ, Feng Z, Adam B-L, Banez LL, Bigbee WL, Campos D, 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;51:102–12.
[94] Grizzle WE, Adam BL, Bigbee WL, Conrads TP, Carroll C, Feng Z, et al. Serum protein expression profiling for cancer detection: validation of a SELDI-based approach for prostate cancer. Dis Markers 2003–2004;19:185–95.
[95] Boguski MS, McIntosh MW. Biomedical informatics for proteomics. Nature 2003;422:233.
[96] Coombes KR, Fritsche Jr HA, Clarke C, Chen JN, Baggerly KA, Morris JS, et al. Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. Clin Chem 2003;49:1615–23.
[97] Feng Z, Prentice R, Srivastava S. Research issues and strategies for genomic and proteomic biomarker discovery and validation: a statistical perspective. Pharmacogenomics 2004;5:709–19.
[98] Hu J, Coombes KR, Morris JS, Baggerly KA. The importance of experimental design in proteomic mass spectrometry experiments: some cautionary tales. Brief Funct Genomics Proteomics 2005;3:322–31.
[99] Aebersold R, Goodlett DR. Mass spectrometry in proteomics. Chem Rev 2001;101:269–95.

[100] Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature 2003;422:198–207.
[101] Gygi SP, Aebersold R. Mass spectrometry and proteomics. Curr Opin Chem Biol 2000;4:489–94.
[102] Mann M, Hendrickson RC, Pandey A. Analysis of proteins and protemes by mass spectrometry. Annu Rev Biochem 2001;70:437–73.
[103] Siuzdak G. Mass spectrometry for biotechnology. San Diego, CA: Academic Press; 1996.
[104] Yates 3rd JR. Mass spectrometry. From genomics to proteomics. Trends Genet 2000;16:5–8.
[105] Keller BO, Li L. Discerning matrix-cluster peaks in matrix-assisted laser desorption/ionization time-of-flight mass spectra of dilute peptide mixtures. J Am Soc Mass Spectrom 2000;11:88–93.
[106] Krutchinsky AN, Chait BT. On the nature of the chemical noise in MALDI mass spectra. J Am Soc Mass Spectrom 2002;13:129–34.
[107] Hutchens TW, Yip TT. New desorption strategies for the mass spectrometric analysis of macromolecules. Rapid Commun Mass Spectrom 1993;7:576–80.
[108] Issaq HJ, Veenstra TD, Conrads TP, Felschow D. The SELDI-TOF MS approach to proteomics: protein profiling and biomarker identification. Biochem Biophys Res Commun 2002;292:587–92.
[109] Merchant M, Weinberger SR. Recent advancements in surface-enhanced laser desorption/ionization-time of flight-mass spectrometry. Electrophoresis 1999;21:1164–677.
[110] Tang N, Tornatore P, Weinberger SR. Current developments in SELDI affinity technology. Mass spectrom Rev 2004;1:34–44.
[111] Wu W, Hu W, Kavanagh JJ. Proteomics in cancer research. Intl J Gynecol Cancer 2002;12:409–23.
[112] Vander A, Sherman J, Luciano D. Human physiology. 8th ed. New York: McGraw-Hill; 2001.
[113] Anderson NL, Anderson NG. The human plasma proteome: history, character, and diagnostic prospects [erratum appears in mol cell proteomics. 2003 jan;2(1):50]. Mol Cell Proteomics 2002;1:845–67.
[114] Coombes KR, Tsavachidis S, Morris JS, Baggerly KA, Hung MC, Kuerer HM. Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. Houston: M.D. Anderson Cancer Center; 2004.
[115] Satten GA, Datta S, Moura H, Woolfitt AR, Carvalho MdG, Carlone GM, et al. Standardization and denoising algorithms for mass spectra to classify whole-organism bacterial specimens. Bioinformatics 2004;20:3128–36.
[116] Malyarenko DI, Cooke WE, Adam B-L, Malik G, Chen H, Tracy ER, et al. Enhancement of sensitivity and resolution of surface-enhanced laser desorption/ionization time-of-flight mass spectrometric records for serum peptides using time-series analysis techniques. Clin Chem 2005;51:65–74.
[117] Shin H, Koomen J, Baggerly KA, Markey MK. Towards a noise model of MALDI TOF spectra. In: American Association for Cancer Research (AACR) advances in proteomics in cancer research, 2004. Key Biscayne, FL; 2004.
[118] Preparata FR, Shamos MI. Computational geometry: An introduction. New York: Springer; 1985.
[119] Wang MZ, Howard B, Campa MJ, Patz Jr EF, Fitzgerald MC. Analysis of human serum proteins by liquid phase isoelectric focusing and matrix-assisted laser desorption/ionization-mass spectrometry. Proteomics 2003;3:1661–6.
[120] Anderle M, Roy S, Lin H, Becker C, Joho K. Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography–mass spectrometry of human serum. Bioinformatics 2004;446.
[121] Barclay VJ. Application of wavelet transforms to experimental spectra: Smoothing, denoising, and data set compression. Anal Chem 1997:78–90.
[122] Shao X-G, Leung AK-M, Chau F-T. Wavelet: a new trend in chemistry. Acc Chem Res 2003;36:276–83.
[123] Robinson EA. Statistical communication and detection. London: Griffin; 1967.
[124] Kuerer H, Coombes K, Chen JX, Clarke L, Fritsche C, Krishnamurthy H, et al. Association between ductal fluid proteomic expression profiles and the presence of lymph node metastases in women with breast cancer. Surgery 2004;136:1061–9.
[125] Prados J, Kalousis A, Sanchez JC, Allard L, Carrette O, Hilario M. Mining mass spectra for diagnosis and biomarker discovery of cerebral accidents. Proteomics 2004;4:2320–32.
[126] Duda RO, Hart PE, Stork DG. Pattern classification. 2nd ed. New York: Wiley-Interscience; 2000.
[127] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157–82.
[128] Fung ET, Enderwick C. Proteinchip clinical proteomics: computational challenges and solutions. Biotechniques 2002(Suppl.).
[129] Jain AK, Duin RPW, Jianchang M. Statistical pattern recognition: A review. IEEE Trans Pattern Anal Mach Intell 2000;22:4.
[130] Blum AL, Langley P. Selection of relevant features and examples in machine learning. Artif Intell 1997;97:245–71.
[131] Dash M, Liu H. Feature selection for classification. Intell Data Anal 1997;1.
[132] Hall MA, Holmes G. Benchmarking attribute selection techniques for discrete class data mining. IEEE Trans Knowl Data Eng 2003;15:1437–47.
[133] Vlahou A, Schellhammer PF, Mendrinos S, Patel K, Kondylis FI, Gong L, et al. Development of a novel proteomic approach for the detection of transitional cell carcinoma of the bladder in urine. Am J Pathol 2001;158:1491–502.
[134] Mitchell TM. Machine learning. Boston: WCB/McGraw-Hill; 1997.
[135] Metz CE. Roc methodology in radiologic imaging. Invest Radiol 1986;21:720–33.
[136] Lancashire LJ, Mian S, Ellis IO, Rees RC, Ball GR. Current developments in the analysis of proteomic data: artificial neural network data mining techniques for the identification of proteomic biomarkers related to breast cancer. Curr Proteomics 2005;2:15–29.
[137] Cristianini N, Scholkopf B. Support vector machines and kernel methods: the new generation of learning machines. AI Magazine 2002.
[138] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. 1st ed. Cambridge: Cambridge University Press; 2000.
[139] Pontil M, Verri A. Properties of support vector machines. Neural Comput 1998;10:955–74.
[140] Bauer E, Kohavi R. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Mach Learn 1999;36:105–39.
[141] Breiman L. Random forests. Mach Learn 2001;45:5–32.
[142] Japkowicz N. Learning from imbalanced data sets: a comparison of various strategies. Technical Report. Menlo Park, CA: AAAI Press; 2000. Report No.: WS-00-05.
[143] Kotsiantis SB, Pintelas PE. Mixture of expert agents for handling imbalanced data sets. Ann Math Comput Teleinformatics 2003;1:46–55.
[144] Maloof MA. Learning when data sets are imbalanced and when costs are unequal and unknown. Washington, DC: Department of Computer Science, Georgetown University; 2003.
[145] Maletic JI, Marcus A. Data cleansing: beyond integrity analysis. In: Information quality (IQ2000); 2000 October 2000. Boston, MA; 2000. p. 200–9.

[146] Orr K. Data quality and systems theory. Commun ACM 1998:66–71.
[147] Redman TC. The impact of poor data quality on the typical enterprise. Commun ACM 1998:79–82.
[148] Dodd LE, Wagner RF, Armato 3rd ed SG, McNitt-Gray MF, Beiden S, Chan HP, et al. Assessment methodologies and statistical issues for computer-aided diagnosis of lung nodules in computed tomography: contemporary research topics relevant to the lung image database consortium. Acad Radiol 2004;11:462–75.
[149] Han J, Kamber M. Data mining: concepts and techniques. 1st ed. San Diego: Academic Press; 2001.
[150] Brodley CE, Friedl MA. Identifying and eliminating mislabeled training instances. In: The 13th national conference on artificial intelligence. Portland, OR: AAAI Press; 1996. p. 799–805.
[151] Gamberger D, Lavrac N, Groselj C. Experiments with noise filtering in a medical domain. In: International conference of machine learning (ICML'99). San Francisco, CA: Morgan Kaufmann; 1999. p. 143–51.
[152] Efron B, Gong G. A leisurely look at the bootstrap, the jackknife, and cross-validation. Am Statistician 1983;37:36–48.
[153] Efron B, Tibshirani R. Statistical data analysis in the computer age. Science 1991;253.
[154] Efron B, Tibshirani RJ. An introduction to the bootstrap. New York, NY: Chapman & Hall; 1993.
[155] Bishop CM. Neural networks for pattern recognition. New York: Oxford University Press; 1995.
[156] Fukunaga K. Introduction to statistical pattern recognition. 2nd ed. Boston: Academic Press; 1990.
[157] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Berlin: Springer; 2002.