DDT: BIOSILICO Vol. 2, No. 5 September 2004

Data mining for mass spectra-based cancer diagnosis and biomarker discovery

Melanie Hilario, Alexandros Kalousis, Julien Prados and Pierre-Alain Binz

Rapid advances in mass spectrometry have positioned it as a prime tool for diagnosis and biomarker discovery. Distilling a handful of accurate disease markers from the thousands of mass-to-charge ratios that comprise a mass spectrum raises non-trivial data analytic challenges. Thus, mass-spectral analysis is turning more and more to data mining, a technology that lies at the crossroads of artificial intelligence and statistical data analysis. This review describes recent attempts at applying data mining techniques to extract diagnostic biomarkers for cancer from SELDI-TOF and MALDI-TOF mass spectra.

Melanie Hilario*
Alexandros Kalousis
Julien Prados
Artificial Intelligence Laboratory
CSD, University of Geneva
24 rue Général-Dufour
1211 Geneva 4
Switzerland
*e-mail:
melanie.hilario@cui.unige.ch
Pierre-Alain Binz
Swiss Institute of Bioinformatics
1 rue Michel Sevret
CH-1211 Geneva
Switzerland

▼ A distinctive feature of the post-genomic era is the increasing focus on the proteome, the ensemble of protein forms expressed in a biological sample at a given point in time. Unlike the genome, the proteome reflects both the intrinsic genetic program of the cell and the impact of its immediate environment; it is a snapshot of the actual functional state of specific cells, tissues or organs. Clinical proteomics aims to study changes in protein expression to discover new disease markers and drug targets. Among its enabling technologies, mass spectrometry (MS) is emerging as a key tool for biomarker discovery [1]. Mass spectrometers have high throughput and resolution; state-of-the-art instruments now cover a wide range of molecular weights in small biological specimens, making mass spectra-based protein analysis possible. Body fluids such as serum or urine have been used to generate protein profiles [mass-to-charge (m/z) ratios versus signal intensities] replete with potential disease markers [2,3]. These biomarker patterns are not always obvious to the human eye; to discover and interpret them, biologists have traditionally used standard statistical methods and are now turning more and more to data mining.

Data mining meets clinical proteomics

According to a widely accepted definition, knowledge discovery in databases (KDD), more familiarly known as data mining, is a non-trivial process of identifying valid, novel, potentially useful and, ultimately, understandable patterns in data [4]. Figure 1 gives an overview of the KDD process. As shown in the figure, the heart of the process is the learning step or the automatic construction of a predictive model by generalizing from the training data. A few machine learning techniques used for this step are described briefly in Box 1.

Mass-spectral data have several distinctive characteristics that should guide a data miner's technological choices throughout the KDD process. First, the quality of a mass spectrum depends on several factors, which vary widely across experiments; hence the need for meticulous data preprocessing and quality assessment before biomarker extraction.

Second, a sample usually contains thousands of different m/z ratios, each with a corresponding signal intensity. These ratios are the variables; their number determines problem dimensionality. Each case can thus be viewed geometrically as a point in a very high-dimensional space. In clinical proteomics, the problem of high dimensionality is compounded by small sample size – diseased specimens are relatively rare and difficult to collect, especially if invasive procedures are involved. This twofold pathology, called the high-dimensionality–small-sample problem, is the hallmark of microarray and mass-spectral data and presents major technological challenges that data miners are only just beginning to address [5,6,7].

Finally, biomarker discovery is the selection of a small panel of proteins from the thousands

 PII: S1741-8364(04)02416-3

of m/z points in mass spectra. Variable selection is not merely a preprocessing expedient aimed at dimension-reduction but is an intrinsic part of the learning task. A corollary requirement is that of model interpretability: the selected variables, and their respective roles and interactions must not only be accessible in the final classifier, they must also make biological sense.

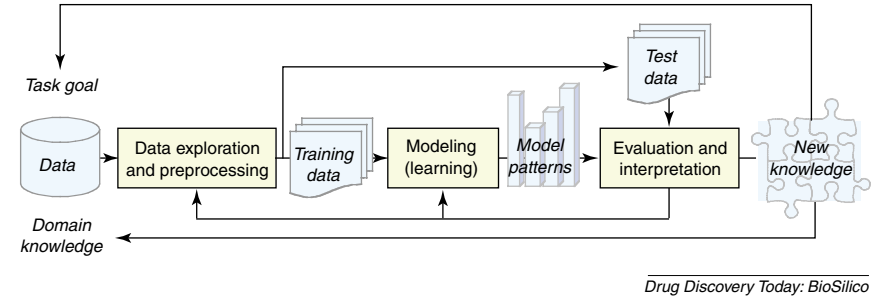

Figure 1. The KDD or data mining process. Data preprocessing involves data cleaning and transformation into a representation that is appropriate for learning (e.g. for mass spectra: denoising, peak detection). Part of the data is then used to build a classification model, while the rest is reserved for testing; depending on model evaluation results, one can reenter the KDD loop or interpret the model for integration into existing domain knowledge.

Preprocessing mass spectra for knowledge discovery

The goal of preprocessing is to transform mass spectra, as acquired by biologists, into a representation that is appropriate for data mining. The main preprocessing steps are baseline and noise elimination, peak detection, mass alignment and intensity normalization. Within this preprocessing, 'baseline correction' involves the extraction of low frequencies from the spectrum signal. The baseline is usually estimated within a local neighborhood, using for example, convolution [8], Savitzky-Golay [9] or mathematical morphology-based filters [10]. The process of 'denoising' is also performed; irrelevant signals such as chemical or electronic noise are eliminated, by means of wavelet transform filters [11,12], as well as filters that are used for baseline removal. Peak detection can be achieved by simply taking local maxima [13] or via wavelets [11], peak models [14,10] or unified approaches that solve the baseline, noise and peak detection problems together [15]. Mass-alignment is aimed at correcting intersample shifts in m/z values that

Box 1. A sampling of machine learning techniques

1. Decision trees and rules express logical combinations of constraints on selected variables. A decision rule is typically a conjunction of several conditions: IF *cond1* AND *cond2* AND ... *condN* THEN *conclusion*. In a decision tree, a rule is simply a path from the root to a terminal node. The tree itself is a disjunction over all these rules (paths). Most decision trees, such as CART [42] and C4.5 [43], examine a single variable at a time.
2. Naïve Bayes computes the posterior probability of each candidate class, using Bayes' theorem and the simplifying hypothesis that all variables are mutually independent. A new case is assigned to the class with the highest posterior probability. (For Bayesian decision theory and the Naïve Bayes rule, see [44], Chapter 2: Bayesian decision theory, pp. 20–65).
3. In K-nearest-neighbors, no internal model is built; learning is simply storing the training cases. To classify a new case, its K nearest neighbors are identified using a similarity metric, such as Euclidean distance. The new case is assigned to the most frequent class among these K neighbors ([44], Chapter 4: Nonparametric techniques, pp. 177–192).
4. The perceptron builds a linear boundary or hyperplane to separate two classes by computing a linear combination of its inputs ([45], Chapter 3: Single-layer networks, pp. 98–105). Learning occurs in finding the appropriate weights, so that the resulting hyperplane optimally separates the classes. The multilayer perceptron is an extension of the perceptron, which allows for additional layers of hidden units between the input vector and the output classes. It can thus build arbitrary nonlinear decision surfaces and solve more complex classification problems. The standard reference in neural networks for pattern classification is ([45], Chapter 4: The multi-layer perceptron, pp. 116–163).
5. SVMs are based on the principle of structural risk minimization, which governs the trade-off between empirical training error and model complexity. For two-class problems, it has been shown that the optimal trade-off is achieved by a hyperplane that produces the maximal margin of separation between the classes. This hyperplane is found by solving a constrained quadratic optimization problem; the solution can then be expressed exclusively in terms of the data points that lie on the margin, the so-called 'support vectors'. This technique can be applied even if the data are nonlinearly separable; the basic idea is to transform the data via a nonlinear mapping onto a higher-dimensional feature space, where they become linearly separable ([46], Chapter 6: Support vector machines).
6. Ensemble methods build classifiers composed of multiple models, the decisions of which are aggregated to classify a new case. RandomForest is an algorithm that grows ensembles of decision trees by randomly selecting, at each node, a small set of variables on which to split the data [47]. Boosting [48] focuses the learning process on the more difficult cases by iteratively reweighting the training samples.

Table 1. Combinations of variable selection methods and learning algorithms used in different diagnostic and biomarker discovery experiments

Variable selection	LDA	QDA	NB	KNN	C4.5	CART	BDS	RF	MLP	SVM	Cancer
I. Sequential variable selection and learning											
t-statistic	x	x	x	x	x	x		x		x	Ovarian [22]
				x						x	Ovarian [25]
χ^2-statistic			x	x	x				x	x	Ovarian [22]
											Renal [26]
F-ratio	x	x		x	x					x	Lung [41]
Wilcoxon	x										Ovarian [23]
AUC						x	x				Prostate [16]
											Prostate [30]
Mutual info			x		x	x			x		Lung [28]
Relief-F			x		x	x			x		Lung [28]
RandomForest	x	x			x		x		x		Ovarian [25]
SFS	x										Prostate [32]
CFS			x		x	x				x	Ovarian [22]
II. Integrated variable selection and learning											
Boosted univariate linear discriminant											Prostate [13]
Multistage neural network training with nested variable selection											Astrocytoma [35]
Genetic algorithm and self-organizing map											Ovarian [36]
Genetic algorithm and linear discriminant											Lung [21]
Partial least squares projection to latent structure											Lung [39]
Weighted flexible compound covariate method											Lung [19]

Abbreviations: BDS, boosted decision stump; KNN, K-nearest-neighbors; LDA, linear discriminant analysis; MLP, multilayer perceptron; NB, Na ve Bayes; QDA, quadratic discriminant analysis; RF, RandomForest

have occurred as a result of instrumental measurement inaccuracy. Most of the proposed methods [16,17,18] can be optimized in what is ultimately a hierarchical clustering procedure. An innovative approach selects optimal clusters via a genetic algorithm search with a multi-objective fitness function – to maximize the number of peaks from different samples and minimize the number of peaks from the same sample [19]. Intensity normalization is the scaling of peak intensities across different spectra to allow for meaningful comparison. If all spectra contain a protein of known concentration, they can be normalized relative to its peak intensities [20]; otherwise, intensities can be normalized on the total ion current (the summed intensities over all time points) [21,22] or reduced to a boolean indicator of peak presence or absence [13].

Before the above transformations, there is an increasingly perceived need for stringent data quality assessment and control. Low-quality data not only impede knowledge discovery, they can actually mislead data mining algorithms into discovering patterns in noise. Recent studies have suggested that patterns that are extracted from mass spectra might simply reflect structure in noise or experimental artefacts (sample handling, changes in instruments or laboratory protocol) rather than the underlying biology [23,24].

Diagnosis and biomarker discovery

Although mass spectra preprocessing drastically reduces the number of variables, further dimension-reduction is usually required, for two reasons. First, the high-dimensionality–small-sample problem, discussed above, despite progress in the search for algorithms that generalize well in sparse, high-dimensional spaces [7]. Second, in biomarker discovery, variable selection is as important a learning goal as accurate classification. These two goals are generally tackled in sequence, although there have been efforts to weave variable selection more tightly into the modeling process.

Sequential dimensionality-reduction and learning

Here, classifier learning is preceded by a second, more aggressive variable-selection phase; the various variable-selection/learning method combinations are summarized in Table 1.

Individual variable selection

In most supervised learning systems, variable selection makes use of information that is provided by the class labels and is subject to the same methodological precautions as the learning task proper – in particular, a strict separation of training and test samples. All variable selection methods for classification rely on some scoring or ranking function to quantify variable relevance or discriminatory power. The final variable set is selected by setting a threshold on the computed scores or ranks. Many classical statistical tests and measures have been used to determine significant differences in variable importance; examples are the t-statistic [22,25], the F-ratio [22] and the χ^2-statistic [22,26].

Alternative variable-ranking/selection criteria derive from information theory. Mutual information [27] (i.e. information gain) quantifies the information about the class that is gained from observing a specific variable; it is computed as the initial entropy of the class, minus its entropy given the variable. It has been shown to be an effective variable-ranking criterion in lung cancer prediction [28]. Another popular criterion from information technology is the AUC (area under the curve) or area under the 'receiver operating characteristic' (ROC) curve (see Box 2) [29]. The AUC has been used to rank peaks in prostate cancer detection before learning with decision trees [16,30].

All the variable scoring and ranking criteria and methods that have been discussed previously assume mutual independence of all predictor variables; this is often an unrealistic assumption. Machine learning research has given rise to novel methods that take account of variable dependencies. Relief-F [31], for instance, computes the relevance of each predictive variable via a method that is based on K-nearest-neighbors. The underlying idea is that a discriminatory variable should have identical or close values in members of the same class and contrasting values in cases belonging to different classes. Relief-F has proved effective in the presence of interacting variables and has been used for m/z value ranking and selection [28]. A subroutine of the RandomForest [25] learning algorithm can also be used for variable scoring (Box 1, point 6). When a decision tree is built, its misclassification error is estimated on an independent test set. If we randomly permute the values of a given variable among the test samples and reapply the classifier to the noised set, the increase in misclassification rate gives a measure of the importance of the variable. By applying this procedure to all variables, a ranking is obtained that can be used for variable selection.

Variable subset selection

Variable subset selection entails a major difficulty: the number of possible subsets increases exponentially with the number of variables. This precludes exhaustive search for all but trivial datasets, therefore heuristic strategies are needed. Stepwise forward selection (SFS), for example, starts with an empty variable subset, S, and selects the variable that maximizes a predefined scoring function. Thereafter, it selects from the remaining variables that which, added to S, maximizes the score of the resulting subset. The process continues until a predefined criterion is met; for example, until no single variable addition improves the score of the subset. SFS has been used to reduce the peak set in mass-spectral applications [32].

'Correlation-based feature selection' (CFS) [33] is based on the insight that good variable subsets contain variables that are highly correlated with the class, yet are uncorrelated with each other. The merit of a feature subset in CFS is directly proportional to the mean correlation between the subset variables and the class, and inversely proportional to the mean correlation among the variables themselves. Correlation between variables is measured in terms of their symmetrical uncertainty, a normalized form of mutual information. CFS uses stepwise forward selection to find a variable subset that maximizes the merit criterion. It has been found to yield best performance in a comparative study of variable selection methods for ovarian cancer diagnosis [22].

Variable construction

The techniques that have been discussed in the two previous subsections reduce data dimensionality by selecting

Box 2. Model evaluation metrics

Accuracy is the best known measure of classification performance. It is the number of correctly classified examples over the total number of examples in a given dataset. When class distribution is imbalanced, accuracy can be misleading because it is dominated by performance on the majority class. In two-class problems, accuracy can be replaced by sensitivity and/or specificity.

Sensitivity or 'true positive rate' is the number of correctly predicted positive instances over all positive instances. It is the criterion of choice when false negatives incur high penalty, as in most medical diagnosis.

Specificity or 'true negative rate' is the number of correctly predicted negative instances over all negative instances. It is used when false alarms are costly.

An **ROC curve** plots the sensitivity of a classifier against 1 minus its specificity of a classifier, as a threshold on some continuous output (e.g. probability of positive class) is varied. The resulting curve is a 2D representation of classifier performance. When a single global measure is needed, the ROC AUC can be computed; a higher AUC roughly indicates better average performance. Caveat: when ROC curves cross, a classifier with a high AUC will not always perform better than one with a low AUC.

from a preexisting set of variables. An alternative approach is to construct new variables by combining or transforming the old ones. In 'principal components analysis' (PCA), the new variables, called components, are linear combinations of the original variables. Although theoretically there can be as many components as original variables, often a much smaller set of components can explain most of the variability in the data. Substantial dimensionality reduction can thus be attained by describing the data in terms of a few principal components. Lilien and co-workers [34] used PCA to reduce raw m/z ratios (15 000–16 000 variables) in view of linear discriminant analysis. PCA is a classical statistical technique but machine learning research on constructive induction has produced a body of techniques for variable transformation/construction, as well as predicate invention. Results of this research have remained by and large untapped in mass-spectral data analysis.

Integrated variable selection and model building

Nesting variable selection in the learning loop

Yasui and colleagues' [13] boosted linear discriminant for prostate cancer detection captures the variable selection process within the boosting loop (Box 1, point 6). The base classifier is a univariate discriminant for a two-class problem (for example, cancer versus controls). At every iteration, each candidate variable is used to build a logistic regression model and the model or variable that maximizes the likelihood ratio is selected. The linear part of the selected model becomes the base classifier; its predictions on the training set are evaluated and the weights of misclassified cases are increased (and those of correctly classified cases decreased) for the next iteration. Boosting halts when observed sensitivity and specificity exceed predefined thresholds. After the last iteration M, the final classifier can be written as the sum of M univariate discriminants. This method produced aggregate classifiers with around 25 markers.

A neural network framework for multistage biomarker selection has been proposed for discriminating low-grade and high-grade astrocytomas, based on 12 SELDI-TOF (surface-enhanced laser desorption ionization time-of-flight) mass spectra of diseased tissue [35]. The input spectra were split into blocks, each of which was used to train a multilayer perceptron. Weights of trained networks were examined to determine the most important inputs, which were then combined into a single network for a second round of training and analysis. The top 50 m/z values were subjected to SFS, yielding a final set of four variables, three of which allowed tumour grade prediction to an accuracy of 98% on the training set. This experiment was intended as a proof-of-principle and awaits validation.

Genetic algorithm-based variable subset selection

One of the early studies in mass spectra-based ovarian cancer diagnosis blended genetic algorithms and self-organizing maps into an integrated classification and feature-selection system [36]. Aside from baseline subtraction and intensity scaling to *[0..1]*, the initial dataset did not undergo any preprocessing and contained 15 154 distinct m/z values or potential predictors. Genetic algorithms [37] were used to evolve an initial generation of 1500 sets of 5–20 m/z values into a few selected discriminatory patterns. Training samples were represented using each candidate m/z set and clustered into self-organizing maps [38]. The fitness test was the ability of each set to specify a map with homogeneous cancer and control clusters. Variable sets that were deemed fit were used to spawn new sets through crossover and mutation. The learning process halted after 250 generations or when a map was found that perfectly separated the cancer and control cases. The process yielded an ovarian cancer biomarker with five m/z values.

Baggerly and co-workers [21] combined genetic algorithms with linear discriminants to diagnose lung cancer from MALDI-TOF (matrix-assisted laser desorption ionization time-of-flight) spectra of serum samples. Learning was preceded by comprehensive preprocessing and dimension reduction to 506 m/z values. Biomarkers of N=1 to N=5 peaks were generated and tested; exhaustive search was used for N=1 and N=2, and genetic algorithms for N=3, N=4 and N=5. For each N, fifty genetic algorithm experiments were conducted, each with a different initial population of 200 sets of N peaks. Each peak set was used to perform a linear discriminant analysis on the training samples; the sets that maximized the Mahalanobis distance between the cancer and control groups were used to generate new variable sets. (Roughly stated, the Mahalanobis distance between two groups is the distance between their centroids, normalized by their covariance matrix.) Training halted after 250 generations. The experiments led to the identification of a best single peak as well as a five-peak biomarker pattern.

Variable construction for discrimination

Lee and co-workers' system [39] uses a technique that reduces dimensionality, while solving a discrimination problem. 'Partial least squares projection' (PLS) to latent structure can be viewed as the supervised counterpart of principal components analysis. It extracts latent variables as linear combinations of the original explanatory variables, such that most of their association with the response variable is explained. Dimensionality is reduced when the first few linear combinations of predictors explain most of the association with the response. PLS was used as a discriminant

analysis tool (PLS-DA) to separate lung cancer cases and controls from the Duke University dataset (no longer available online). The initial 60 000 m/z values were reduced via a wavelet transform to 545 wavelet coefficients. PLS-DA produced a two-component discriminant model. Each component was a linear combination of wavelet coefficients, which were then inverse-transformed to the original variates to identify discriminatory m/z ratios.

Yanasigawa and colleagues [19] employ six statistical tests, such as the Kruskal-Wallis test and Fisher's exact test, to select a set of m/z points that are differentially expressed between two groups, say lung cancer and normal. They then reduce dimensionality by creating compound covariates (much like Lee and colleagues' latent variables) as the weighted sum of the selected variables. The novelty of their approach, which is called the 'weighted flexible compound covariate method', lies in the incorporation of multiple statistical tests to determine the weights of the original variables in the new covariates.

Model assessment and interpretation

The main evidence of the usefulness of discovered biomarkers is their diagnostic power. In predictive data mining, there are widely accepted metrics (Box 2) and strategies (Box 3) for diagnostic or classification tasks. This section will review performance results on datasets that were accessible to different research teams.

Ovarian cancer datasets

Three datasets are available for ovarian cancer. The first version (Table 2, row 1) gave rise to two studies. Petricoin *et al.*'s [36] machine learning approach extracted a five-marker pattern that achieved 100% sensitivity and 95% specificity on a blind test set of 116 (50 diseased, 66 benign/control) out of a total of 216 samples (i.e. with a 46:54 train:test split). Lilien and colleagues [34] ran their principal components-based linear discriminant, Q5, on the same dataset in a variety of experimental conditions. To ensure a fair comparison with the results of Petricoin's group, we selected parameter settings from Lilien's experiments that were closest to those used by Petricoin: 50:50 train:test split and a probability classification threshold of 0.5, which returned a prediction for 98.04% of the test set (Petricoin's method classified all test samples). Under this setup, Lilien's statistical approach obtained a sensitivity of 87.57% and a specificity of 90.15%. In contrast to the small subset of m/z values harvested by Petricoin, Lilien's model was a linear combination of principal components, which were themselves linear combinations of the original m/z ratios. The discriminant was back-projected onto mass-spectral space to identify a handful of biomarkers among the m/z values with the highest coefficients.

Box 3. Model evaluation strategies

The common rule underlying all classifier evaluation procedures is that training and testing should be conducted on independent or disjoint datasets. When sufficient data are available, the holdout method is the gold standard: data are partitioned once and for all into a training set and a blinded test set. Otherwise, data should be resampled to allow for their reuse, while respecting a strict separation of training and test data. Following are the most popular resampling methods:

K-fold cross-validation: the data are partitioned into K disjoint sets. The classifier is trained on K minus 1 sets, then tested and rated on the remaining set. This is repeated K times, with a different test set at each iteration. The estimated error is simply the average of the K error rates observed.

Leave-one-out cross-validation: a particular case of K-fold cross-validation, where K equals the number of cases in the dataset. Thus, at each iteration, the test set consists of exactly one case.

Bootstrap: given an initial dataset of size N, a new sample of size N is built by randomly selecting cases without removing them from the initial set. This sample is used for training, whereas all cases that are not selected in the sampling process are used for testing. Bootstrapping requires a large number of iterations, typically between 50 and 200.

Three teams experimented with the third ovarian cancer dataset (Table 2, row 3), composed of 253 specimens (162 diseased, 91 controls). All reported 100% sensitivity and 100% specificity as their best results. With a 50% probability classification threshold and a training proportion of at least 75%, Q5 classified all test samples with perfect accuracy [34]. Sorace *et al.* [23] used Wilcoxon variable ranking, followed by stepwise discriminant analysis to train three linear models on 49% of the dataset, then tested these on the remaining data. Two models with different sets of seven m/z values each achieved perfect classification. Liu and colleagues [22] used tenfold cross-validation to compare different combinations of variable selection and learning methods (Table 1). Of the variable selection methods, CFS consistently achieved best performance for all learning algorithms used. Of the learning algorithms, SVM (support vector machines) scored the lowest average error over the different variable selection methods used. Perfect accuracy was achieved by coupling CFS with SVM and with K-nearest-neighbors. Note that SVM achieved perfect classification even without variable selection. Although such a result would appear suspect in the case of many traditional learning methods, it is plausible for SVMs, which have been shown to be robust to the curse of dimensionality and to generalize well in sparse, high-dimensional spaces [7]. A minor caveat: tenfold cross-validation produces more optimistic error estimations than the blind test method as used in the two other studies.

Table 2. Datasets used in comparable biomarker discovery studies

	Disease	MS	Source	Data	Used in
1	Ovarian cancer OC-H4 FDA CPPD	SELDI	serum	216 specimens 100 cancer 100 controls 16 benign	[36,34]
2	Ovarian cancer OC-WCX2a FDA CPPD	SELDI	serum		[34]
3	Ovarian cancer OC-WCX2b FDA Clinical Proteomics Program Databank	SELDI	serum	253 specimens 162 cancer 28 stage I 20 stage II 99 stage III 12 stage IV 3 stage ? 91 controls	[34,22,23]
4	Prostate cancer PC-IMAC-CU E. Virginia Medical School	SELDI	serum	386 specimens 197 PC 99 early-stage 98 late-stage 93 BPH 96 controls	[34,16,30,32,13]
5	Astrocytoma	SELDI	tissue	12 specimens 5 low-grade 7 high-grade	[35]
6	Renal cancer	SELDI	urine	138 specimens	[26]
7	Ovarian cancer Northwestern Univ. Hospital	MALDI-TOF	serum	89 specimens 47 cancer 42 controls	[36]
8	Lung cancer Duke Univ.	MALDI-TOF	serum	41 specimens 24 cancer (A) 17 controls (B)	[21,41,39,28]

Wu and co-workers' [25] comparative study of methods for ovarian cancer diagnosis was based on MALDI-TOF spectra (Table 2, row 8). Two variable selection algorithms – t-statistic-based ranking and RandomForest scoring – were explored in conjunction with six learning methods: linear and quadratic discriminant analysis, K-nearest-neighbors, SVM, boosted CART ('classification and regression trees') and RandomForest. Classification accuracy was estimated using both tenfold cross-validation and bootstrap for the simple algorithms and a 2:1 train:test split for the combined models. Only variable subsets of size 15 and 25 were considered. On both sizes, SVM achieved best performance when variables were ranked according to the t-statistic; however, with RandomForest-based variable selection, RandomForest and boosted CART performed better than SVM. Overall, RandomForest-based variable selection not only led to higher accuracy, it also proved to be more stable than t-statistic-based variable ranking.

Prostate cancer dataset

We analyzed five studies that were based on the prostate cancer dataset. Three of these used the data as described in Table 2, row 4 (two others, [32,34], used different subsets of the data and so are not comparable). All three followed the same 85–15% decomposition of the 386-specimen dataset into training and test sets. Two employed an AUC-based variable-ranking method to reduce the set of candidate markers to 779 m/z values. Adam *et al.* [16] used CART to produce a nine-node decision tree. Qu and colleagues [30] created two ensembles of decision stumps (one-node decision trees), which classified cases via a weighted majority vote. Adaboost (Box 1, point 6) generated a classifier, comprising 500 base classifiers and 74 peaks. A variant called 'boosted decision stump feature selection' (BDSFS), which added the requirement that each variable be used just once or not at all, produced an ensemble of 21 base classifiers. Yasui *et al.*'s [13] approach, which combines

marker selection with linear discriminant analysis within the boosting cycle, required two linear classifiers for this three-class problem: one to distinguish prostate cancer/benign prostatic hyperplasia (PC/BPH) (26 peaks) versus controls (C), a second to separate PC from BPH (25 peaks). We follow Yasui's decomposition to compare the three solutions to the prostate cancer problem in Table 3.

Table 3. Performance measures on the prostate cancer problem

Trained model	PC/BPH vs C		PC vs BPH	
	sensit	specif	sensit	specif
AUC + CART	91	100	83	93
AUC + Adaboost	100	100	100	100
AUC + BDSFS	100	93	97	100
Boosted linear discriminant	98	100	93	47

The results show a classical trade-off in machine learning: the most accurate classifier, built by Adaboost, is also the most difficult to interpret, given its 500 nodes and 74 peaks. Performance degrades slightly with BDSFS (21 nodes/peaks) and, more remarkably, with CART's highly readable tree (nine nodes/peaks). The boosted linear discriminant with integrated marker selection fares reasonably well in discriminating PC/BPH from controls, but is not efficient in distinguishing benign from malignant cancer. Overall, results of the boosted linear discriminant should be considered with caution: the sensitivity and specificity rates that were used to halt the boosting process seem to have been computed on the training data rather on a separate tuning/validation set.

Lung cancer dataset

The lung cancer dataset was the object of a data mining challenge [40] that elicited more than a dozen experimental studies. Three solutions were selected for this review, with the aim of illustrating the diversity of approaches explored. Lee and co-workers [39] built several PLS discriminant models, each with a different experimental strategy. A first model, built on the complete data, achieved 100% accuracy – a result that is both unsurprising and unreliable, as the model was trained and tested on the same data. The dataset was then partitioned into a design set of 28 cases and a test set of 13. A PLS discriminant was built from the design set by sevenfold cross-validation. The resulting two-component model produced one false positive and one false negative, yielding a sensitivity of 87.5% and a specificity of 80%, or an overall accuracy of 85%. The same process, using leave-one-out cross-validation, led to a final two-component model with an accuracy of 76%. The differences in these three accuracy rates illustrate the impact of the error estimation strategy on model assessment and selection.

Wagner *et al.* [41] used F-ratio-based variable ranking, followed by a comparative study of five learning algorithms – linear and quadratic discriminant analysis, kernel-based density estimation, K-nearest neighbors, and SVM. They tested two different experimental protocols to select 3–15 peaks: the first used the full dataset to rank variables before cross-validation, whereas the second integrated variable selection into the leave-one-out cross-validation loop. We will ignore the results of the first strategy, which has the methodological flaw of using test sample labels in variable selection. The second strategy produced its best results with 13-peak models. Linear SVM outperformed all other classifiers, with an accuracy of 98% (96% sensitivity, 100% specificity) by comparison to 73% for the closest runner-up.

Baggerly and colleagues [21] built biomarker patterns of 1–5 peaks using a genetic algorithm–linear discriminant hybrid. The accuracy of these peak sets was then estimated via leave-one-out cross-validation. The best single peak, which appeared in all the best one- to five-peak sets, scored 74%; accuracy increased with peak size, the best five-peak set attaining 98%. Again, these results should be taken with caution; because peak subset selection involving a supervised learning technique (linear discriminant analysis) was performed on the full data before cross-validation, the accuracy rates reported are likely to be optimistic.

From a data mining perspective, the deliverables of diagnosis and biomarker discovery systems are the learned classifier and the proteomic patterns that underly its decisions. This is one reason why most of the studies surveyed convey little about biomarker interpretation. The second reason is that interpreting the biological significance of the extracted proteomic patterns is a long and complex process that involves mapping the selected m/z ratios to proteins (e.g. via database searches or further MS-based experiments), and investigating the intricate biological pathways that link the identified proteins to the disease process. At this final and decisive stage of the biomarker discovery process, the data miner must hand the reins over to the biomedical researcher.

Conclusion

Mass spectrometry is emerging as a gold standard tool for biomarker discovery. It is now possible to envisage early cancer detection through large-scale, non-invasive collection of biological samples, which are then fed into mass spectrometers to generate protein profiles. Analysis of these profiles is relying increasingly on strategies and tools that are provided by data mining research.

However, mass-spectral analysis raises specific data mining challenges. Building accurate models from extremely noisy, high-dimensional and often insufficient data is a task that demands solutions at all stages of the knowledge discovery process. Mass-spectra preprocessing and quality control require close interaction among biomedical and signal processing specialists. Dimensionality reduction has been addressed mainly through the use of standard statistical techniques; many variable-selection/construction techniques from machine learning remain untapped. As for the learning phase, the biggest challenge lies in developing algorithms that generalize well in sparse, high-dimensional spaces.

Knowledge discovery from mass spectra is still in its infancy; the full potential of many techniques for data engineering and machine learning is yet to be exploited in the search for diagnostic biomarkers. This can only be achieved through effective interdisciplinary collaboration that integrates the diagnostic and therapeutic goals with sample collection and preparation constraints, as well as data mining requisites, resources and pitfalls.

References

1 Binz, P.A. *et al.* (2003) Mass spectrometry-based proteomics: current status and potential use in clinical chemistry. *Clin. Chem. Lab. Med.* 41, 1540–1551
2 Wulfkuhle, J. *et al.* (2003) Proteomic applications for the early detection of cancer. *Nat. Rev.* 3, 267–276
3 Srinivas, P.R. *et al.* (2002) Proteomics for cancer biomarker discovery. *Clin. Chem.* 48, 1160–1169
4 Fayyad, U. *et al.* (1996) From data mining to knowledge discovery: an overview. In *Advances in Knowledge Discovery and Data Mining*, MIT Press, pp. 1–34
5 Somorjai, R. *et al.* (2003) Class prediction and discovery using gene microarray and proteomics mass spectrometry data : curses, caveats, cautions. *Bioinformatics* 19, 1484–1491
6 Simon, R. (2003) Supervised analysis when the number of candidate features (p) greatly exceeds the number of cases (n). *SIGKDD Explorations* 5, 31–36
7 Duin, R. (2000) Classifiers in almost empty spaces. In *Proc. 15th Int. Conf. Pattern Recognitio*n, IEEE Computer Society Press, Los Alamitos, vol. 2, pp. 1–7
8 Carroll, J.A. and Beavis, R.C. (1996) Using matrix convolution filters to extract information from time-of-flight mass spectra. *Rapid Commun. Mass Spectrom.* 10, 1683–1687
9 Savitzky, A. and Golay, M. (1964) Smoothing and differentiation of data by simplified least squares procedure. *Anal. Chem.* 36, 1627–1639
10 Breen, E.J. (2000) Automatic Poisson peak harvesting for high throughput protein identification. *Electrophoresis* 21, 2243–2251
11 Shao, X.G. *et al.* (2003) Wavelet: A new trend in chemistry. *Acc. Chem. Res.* 36, 276–283
12 Barclay, V.J. (1997) Application of wavelet transforms to experimental spectra: smoothing, denoising, and data set compression. *Anal. Chem.* 69, 78–90
13 Yasui, Y. *et al.* (2003) A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics* 4, 449–463
14 Gras, R. *et al.* (1999) Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. *Electrophoresis* 20, 3535–3550
15 Mohammad-Djafari, A. *et al.* (2002) Regularization, maximum entropy and probabilistic methods in mass spectrometry data processing problems. *Int. J. Mass Spectrom.* 215, 175–193
16 Adam, B.L. (2002) Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res.* 62, 3609–3614
17 Slotta, D.J. (2003) Clustering mass spectrometry data using order statistics. *Proteomics* 3, 1667–1672
18 Wang, M.Z. (2003) Analysis of human serum proteins by liquid phase isoelectric focusing and matrix-assisted laser desorption/ionization-mass spectrometry. *Proteomics* 3, 1661–1666
19 Yanagisawa, K. *et al.* (2003) Proteomic patterns of tumour subsets in non-small-cell lung cancer. *Lancet* 362, 433–439
20 Clarke, W. (2003) Characterization of renal allograft rejection by urinary proteomic analysis. *Ann. Surg.* 237, 660–665
21 Baggerly, K.A. (2003) A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization time of ight proteomics spectra from serum samples. *Proteomics* 3, 1667–1672
22 Liu, H. *et al.* (2002) A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Informatics* 13, 51–60
23 Sorace, J.M. and Zhan, M. (2003) A data review and reassessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics* 4, 24
24 Baggerly, K.A. *et al.* (2004) Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics* 20, 777–785
25 Wu, B. *et al.* (2003) Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics* 19, 1636–1643
26 Rogers, M.A. *et al.* (2003) Proteomic profiling of urinary proteins in renal cancer by surface enhanced laser desorption ionization and neural-network analysis: identification of key issues affecting potential clinical utility. *Cancer Res.* 63, 6971–6983
27 Cover, T. and Thomas, J. (1991) Elements of information theory. *Chapter* 2 Entropy, Relative Entropy and Mutual Information, 12–49
28 Hilario, M. *et al.* (2003) Machine learning approaches to lung cancer prediction from mass spectra. *Proteomics* 3, 1716–1719
29 Pepe, M.S. (2000) Receiver Operating Characteristic methodology. *J. Am. Stat. Assoc.* 95, 308–311
30 Qu, Y. *et al.* (2002) Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clin. Chem.* 48, 1835–1843
31 Kononenko, I. (1994) Estimating attributes: analysis and extensions of RELIEF. In *Proc. European Conference on Machine Learning*, Springer, pp. 171–182
32 Qu, Y. *et al.* (2003) Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensional data. *Biometrics* 59, 143–151
33 Hall, M. and Holmes, G. (2003) Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering*, 15, 1437-–1447
34 Lilien, R.H. *et al.* (2003) Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. *J. Comput. Biol.* 10, 925–946
35 Ball, G. *et al.* (2002) An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. *Bioinformatics* 18, 395–404
36 Petricoin, E.F., III *et al.* (2002) Use of proteomic patterns in serum of identify ovarian cancer. *Lancet* 359, 572–577
37 Holland, J. (1992) *Adaptation in Natural and Aritificial Systems*, MIT Press
38 Kohonen, T. (1995) *Self-Organizing Maps*, Springer-Verlag
39 Lee, K.R. (2003) Megavariate data analysis of mass spectrometric proteomics data using latent variable projection method. *Proteomics* 3, 1680–1686
40 Campa, M. *et al.* (2003) Editorial. *Proteomics* 3, 1659–1660
41 Wagner, M. *et al.* (2003) Protocols for disease classification from mass spectrometry data. *Proteomics* 3, 1692–1698
42 Breiman, L. *et al.* (1984) *Classification and Regression Trees*, Wadsworth, Belmont, CA
43 Quinlan, J.R. (1993) *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA
44 Duda, R. *et al.* (2000) *Pattern Classification*, Wiley
45 Bishop, C.M. (1995) *Neural Networks for Pattern Recognition*, Oxford University Press
46 Cristianini, N. and Shawe-Taylor, J. (2000) *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press.
47 Breiman, L. (2001) Random forests. *Mach. Learn.* 45, 5–32
48 Freund, Y. and Schapire, R.E. (1996) Experiments with a new boosting algorithm. In *Machine Learning: Proc. 13th International Conference*, Morgan Kaufmann, pp. 148–156