34. Statistical Methods in Proteomics

Proteomics technologies are rapidly evolving and attracting great attention in the post-genome era. In this chapter, we review two key applications of proteomics techniques: disease biomarker discovery and protein/peptide identification. For each of the applications, we state the major issues related to statistical modeling and analysis, review related work, discuss their strengths and weaknesses, and point out unsolved problems for future research.

We organize this chapter as follows. Section 34.1 briefly introduces mass spectrometry (MS) and tandem MS/MS with a few sample plots showing the data format. Section 34.2 focuses on MS data preprocessing. We first review approaches in peak identification and then address the problem of peak alignment. After that, we point out unsolved problems and propose a few possible solutions.

Section 34.3 addresses the issue of feature selection. We start with a simple example showing the effect of a large number of features. Then we address the interaction of different features and discuss methods of reducing the influence of noise. We finish this section with some discussion on the application of machine learning methods in feature selection. Section 34.4 addresses the problem of sample classification. We describe the random forest method in detail in Sect. 34.5.

In Sect. 34.6 we address protein/peptide identification. We first review database searching methods in Sect. 34.6.1 and then focus on de novo MS/MS sequencing in Sect. 34.6.2. After reviewing major protein/peptide identification programs like SEQUEST and MASCOT in Sect. 34.6.3, we conclude the section by pointing out some major issues that need to be addressed in protein/peptide identification.

Proteomics technologies are considered the major player in the analysis and understanding of protein function and biological pathways. The development of statistical methods and software for proteomics data analysis will continue to be the focus of proteomics for years to come.

34.1 Overview

In the post-genome era, proteomics has attracted more and more attention due to its ability to probe biological functions and structures at the protein level. Although recent years have witnessed great advancement in the collection and analysis of gene expression microarray data, proteins are in fact the functional units that

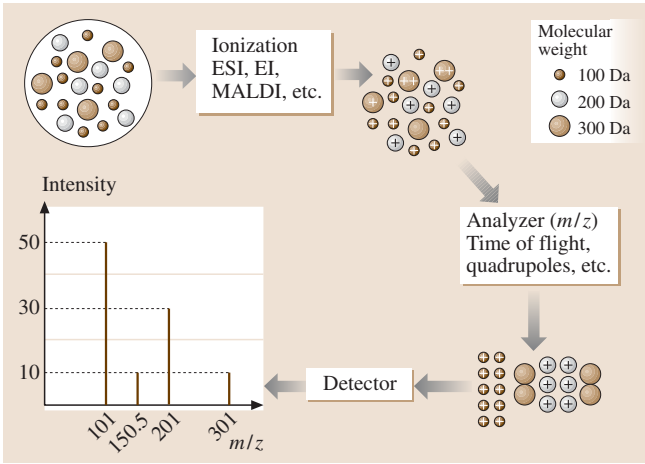

Fig. 34.1 The principle of MS data generation. Molecules are ionized into peptides in the ionizer. The peptides are accelerated and separated by the analyzer and then detected by the detector

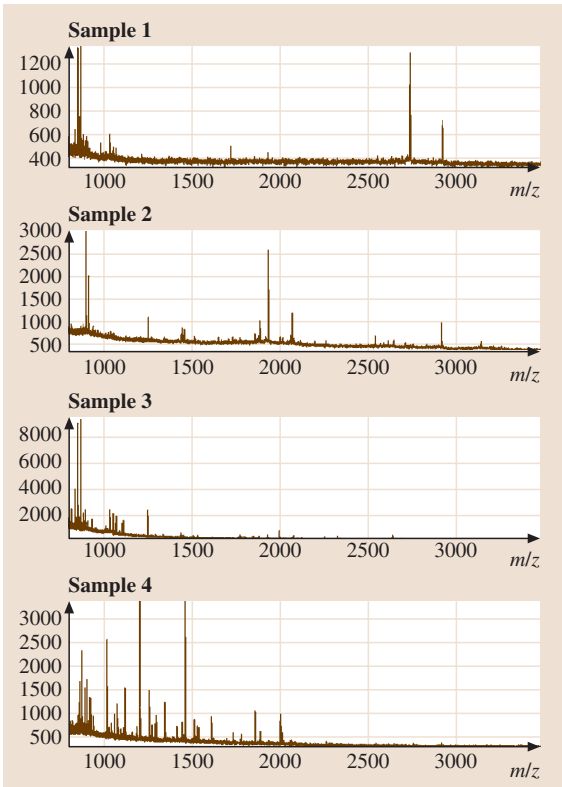

Fig. 34.2 A few examples of MS raw data. The horizontal axis denotes the m/z ratio and the vertical axis denotes the intensity value

are of biological relevance. The often poor correlation that exists between levels of mRNA versus protein expression [34.1], and the rapid advances in mass spectrometry (MS) instrumentation and attendant protein profiling methodologies have substantially increased interest in using MS approaches to identify peptide and protein biomarkers of disease. This great level of interest arises from the high potential of biomarkers to provide earlier diagnosis, more accurate prognosis and disease classification; to guide treatment; and to increase our understanding at the molecular level of a wide range of human diseases. This chapter focuses on two key applications of proteomics technologies: disease biomarker discovery and protein identification through MS data. We anticipate that statistical methods and computer programs will contribute greatly to the discovery of disease biomarkers as well as the identification of proteins and their modification sites. These methods should help biomedical researchers to better realize the potential contribution of rapidly evolving and ever more sophisticated MS technologies and platforms.

The study of large-scale biological systems has become possible thanks to emerging high-throughput mass spectrometers. Basically, a mass spectrometer consists of three components: ion source, mass analyzer, and detector. The ion source ionizes molecules of interest into charged peptides, the mass analyzer accelerates these peptides with an external magnetic field and/or electric field, and the detector generates a measurable signal when it detects the incident ions. This procedure of producing MS data is illustrated in Fig. 34.1. Data resulting from MS sources have a very simple format consisting entirely of paired mass-to-charge ratio (m/z value) versus intensity data points. Figure 34.2 shows a few examples of the raw MALDI-MS data.

The total number of measured data points is extremely large (about 10^5 for a conventional MALDI-TOF instrument, as compared to perhaps 10^6 for a MALDI-FTICR instrument covering the range from 700–28 000 Da), while the sample size is usually on the order of hundreds. This very high ratio of data size to sample size poses unique statistical challenges in MS data analysis. It is desirable to find a limited number of potential peptide/protein biomarkers from the vast amount of data in order to distinguish cases from controls and enable classification of unknown samples. This process is often referred to as biomarker discovery. In this chapter, we review key steps in biomarker discovery: preprocessing, feature selection, and sample classification.

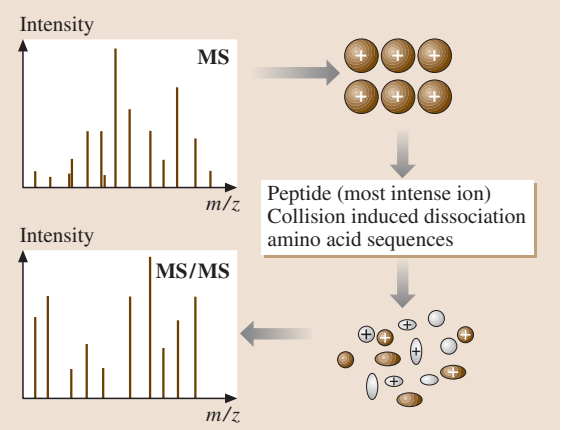

Fig. 34.3 The principle of MS/MS data generation. Peptides are further fragmented through collision-induced dissociation (CID) and detected in the tandem MS equipment

A biomarker discovered in the MS data may correspond to many possible biological sources (so a spectral peak can arise from different proteins). Therefore, it is necessary to identify peptides and their parent proteins in order to fully understand the relation between protein structure and disease development. This understanding can also be very useful in drug design and development.

In order to identify proteins in complex mixtures, the tandem MS technique (MS/MS) coupled with database searching has become the method of choice for the rapid and high-throughput identification, characterization, and quantification of proteins. In general, a protein mixture of interest is enzymatically digested, and the resulting peptides are then further fragmented through collision-induced dissociation (CID). The resulting tandem MS spectrum contains information about the constituent amino acids of the peptides and therefore information about their parent proteins. This process is illustrated in Fig. 34.3.

Many MS/MS-based methods have been developed to identify proteins. The identification of peptides containing mutations and/or modifications, however, is still a challenging problem. Statistical methods need to be developed to improve identification of modified proteins in samples consisting of only a single protein and also in samples consisting of complex protein mixtures.

We organize the rest of the chapter as follows: Section 34.2 describes MS data preprocessing methods. Section 34.3 focuses on feature selection. Section 34.4 reviews general sample classification methodology and Sect. 34.5 mainly describes the random forest algorithm. Section 34.6 surveys different algorithms/methods for protein/peptide identification, each with its strengths and weaknesses. It also points out challenges in the future research and possible statistical approaches to solving these challenges. Section 34.7 summarizes the chapter.

34.2 MS Data Preprocessing

When analyzing MS data, only the spectral peaks that result from the ionization of biomolecules such as peptides and proteins are biologically meaningful and of use in applications. Different data preprocessing methods have been proposed to detect and locate spectral peaks. A commonly used protocol for MS data preprocessing consists of the following steps: spectrum calibration, baseline correction, smoothing, peak identification, intensity normalization and peak alignment [34.2–4].

Preprocessing starts with aligning individual spectra. Even with the use of internal calibration, the maximum observed intensity for an internal calibrant may not occur at exactly the same m/z value in all spectra. This challenge can be addressed by aligning spectra based on the maximum observed intensity of the internal calibrant. For the sample collected, the distance between each pair of consecutive m/z ratios is not constant. Instead, the increment in m/z values is approximately a linear function of the m/z values. Therefore, a log-transformation of m/z values is needed before any analysis is performed so that the scale on the predictor is roughly comparable across the range of all m/z values. In addition to transforming the m/z values, we also need to log-transform intensities to reduce the dynamic range of the intensity values. In summary, log-transformations are needed for both m/z values and intensities as the first step in MS data analysis. Figure 34.4 shows an example of MS data before and after the log-transformation.

Chemical and electronic noise produce background fluctuations, and it is important to remove these background fluctuations before further analysis. Local smoothing methods have been utilized for baseline

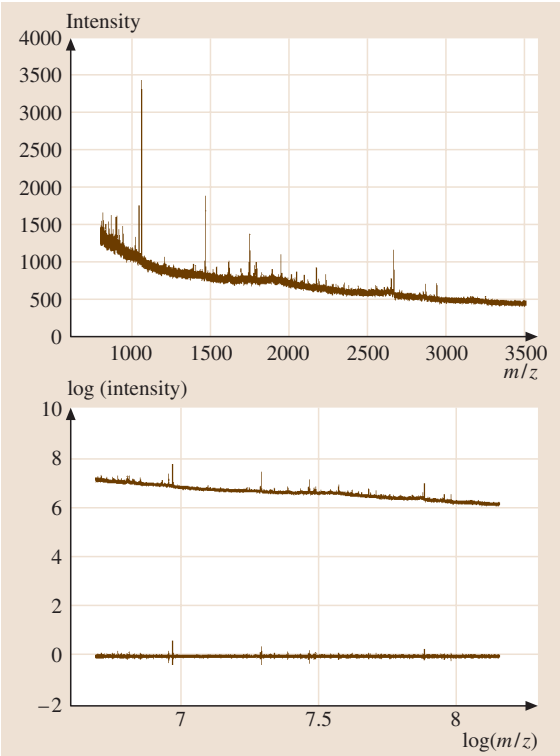

Fig. 34.4 ***Top:*** Original raw data. ***Bottom:*** MS data after the log-transformation (*top*). The result of baseline correction is also shown (*bottom*)

subtraction to remove high frequency noise, which is apparent in MALDI-MS spectra. In the analysis of MALDI data, *Wu* et al. [34.5] used a local linear regression method to estimate the background intensity values, and then subtracted the fitted values from the local linear regression result. *Baggerly* et al. [34.4] proposed a semi-monotonic baseline correction method in their analysis of SELDI data. *Liu* et al. [34.6] computed the convex hull of the spectrum, and subtracted the convex hull from the original spectrum to get the baseline-corrected spectrum. An example of baseline correction is shown in Fig. 34.4 as well.

Among the above steps, peak identification and alignment are arguably the most important ones. The inclusion of non-peaks in the analysis will undoubtedly reduce our ability to identify true biomarkers, while the peaks identified need to be aligned so that the same peptide corresponds to the same peak value.

In the following, we give an overview of the existing approaches related to peak detection and peak alignment.

34.2.1 Peak Detection/Finding

Normally, spectral peaks are local maxima in MS data. Most published algorithms on peak identification use local intensity values to define peaks; in other words peaks are mostly defined with respect to nearby points. For example, *Yasui* et al. [34.3, 7] defined a peak as the m/z value that has the highest intensity value within its neighborhood, where the neighbors are the points within a certain range from the point of interest. In addition, a peak must have an intensity value that is higher than the average intensity level of its broad neighborhood. *Coombes* et al. [34.4] considered two peak identification procedures. For simple peak finding, local maxima are first identified. Then, those local maxima that are likely noise are filtered out, and nearby maxima that likely represent the same peptides are merged. There is a further step needed to remove unlikely peak points. In simultaneous peak detection and baseline correction, peak detection is first used to obtain a preliminary list of peaks, and then the baseline is calculated by excluding candidate peaks. The two steps are iterated and some peaks are further filtered out if the signal-to-noise ratio is smaller than a threshold. Similarly, *Liu* et al. [34.6] declared a point in the spectrum to be a peak if the intensity is a local maximum, its absolute value is larger than a particular threshold, and the intensity is larger than a threshold times the average intensity in the window surrounding this point.

All of these methods are based on similar intuitions and heuristics. Several parameters need to be specified beforehand in these algorithms, such as the number of neighboring points and the intensity threshold value. In fact, the parameter settings in the above algorithms are related to our understanding/modeling of the underlying noise. To address this issue, *Coombes* et al. [34.4] defined noise as the median absolute value of intensity. *Satten* et al. [34.8] used the negative part of the normalized MS data to estimate the variance of the noise. Wavelet-based approaches [34.9, 10] have also been proposed to remove noise in the MS data before peak detection. Based on the observation that there are substantial measurement errors in the measured intensities, *Yasui* and colleagues [34.3] argued that binary peak/non-peak data is more useful than the absolute values of intensity, while they still used a local maximum search method to detect peaks. Clearly, the success of noise-estimation-plus-threshold methodology depends largely on the validity of the noise model, which remains to be seen.

Another issue in peak detection is to avoid false positive detections. This is often done by adding an additional constraint (such as the peak width constraint [34.3]) or by choosing a specific scale level after wavelet decomposition of the original MS data (*Randolph* and *Yasui*) [34.10]. In the case of high-resolution data, it has been proposed that more than one isotopic variant of a peptide peak should be present before a spectral peak is considered to result from peptide ionization (*Yu* et al.) [34.11]. It may also be possible to use prior information about the approximate expected peak intensity distribution of different isotopes arising from the same peptide during peak detection; the theoretical relative abundance of the first peptide isotope peak may range from 60.1% for polyGly (n=23, MW 1329.5 Da) to 90.2% for poly Trp (n=7, MW 1320.5 Da) (personal communication 11/1/04 from Dr. Walter McMurray, Keck Laboratory). Certainly we also have to consider the issue of limited resolution and the consequent overlapping effect of neighboring peaks.

34.2.2 Peak Alignment

After peaks have been detected, we have to align them together before comparing peaks in different data sets. Previous studies have shown that the variation of peak locations in different data sets is nonlinear [34.12, 13]. The example in *Yu* et al. [34.11] shows that this variation still exists even when we use technical replicates. The reasons that underlie data variation are extremely complicated, including differences in sample preparation, chemical noise, cocrystallization and deposition of the matrix-sample onto the MALDI-MS target, laser position on the target, and other factors. Although it is of interest to identify these reasons, we are more interested in finding a framework to reduce the variation and align these peaks together.

Towards this direction, some methods have been proposed. *Coombes* et al. [34.4] pooled the list of detected peaks that differed in location by three clock ticks or by 0.05% of the mass. *Yasui* et al. [34.3] believed that the m/z axis shift of the peaks is approximately $\pm$ 0.1% to $\pm$0.2% of the m/z value. Thus, they expanded each peak to its local neighborhood with the width equal to 0.4% of the m/z value of the middle point. This method certainly oversimplifies the problem. In another study (*Yasui* et al.) [34.7], they first calculated the number of peaks in all samples allowing certain shifts, and selected m/z values using the largest number of peaks. This set of peaks is then removed from all spectra and the procedure is iterated until all peaks are exhausted from all the samples. In a similar spirit, *Tibshirani* et al. [34.14] proposed to use complete linkage hierarchical clustering in one dimension to cluster peaks, and the resulting dendrogram is cut off at a given height. All of the peaks in the same cluster are considered to be the same peak in further analysis.

Randolph and *Yasui* [34.10] used wavelets to represent the MS data in a multiscale framework. They used a coarse-to-fine method to first align peaks at a dominant scale and then refine the alignment of other peaks at a finer scale. From a signal representation point of view, this approach is very interesting. But it remains to be determined whether the multiscale representation is biologically reasonable.

Johnson et al. [34.15] assumed that the peak variation is less than the typical distance between peaks and they used a closest point matching method for peak alignment. The same idea was also used in *Yu* et al. [34.11] to address the alignment of multiple peak sets. Certainly, this method is limited by the data quality and it cannot handle large peak variation.

Dynamic programming (DP) based approaches [34.12, 16] have also been proposed. DP has been used in gene expression analysis to warp one gene expression time series to another similar series obtained from a different biological replicate [34.17], where the correspondence between the two gene expression time series is guaranteed. In MS data analysis, however, the situation is more complicated since a one-to-one correspondence between two data sets does not always exist. Although it is still possible to apply DP to deal with the lack-of-correspondence problem, some modifications are necessary (such as adding an additional distance penalty term to the estimation of correspondence matrix). It also remains unclear how DP can identify and ignore outliers during the matching.

Eilers [34.13] proposed a parametric warping model with polynomial functions or spline functions to align chromatograms. In order to fix warping parameters, he added calibration example sequences into chromatograms. While the idea of using a parametric model is interesting, it is difficult to repeat the same parameter estimation method in MS data since we cannot add many calibrator compounds into the MS samples. Also, it is unclear whether a second-order polynomial would be enough to describe the nonlinear shift in the MS peaks.

Although all of these methods are ad hoc, the relatively small number of peaks (compared to the number of collected points) and the relatively small shifts from spectrum to spectrum ensure that these heuristic peak alignments should work reasonably well in practice.

34.2.3 Remaining Problems and Proposed Solutions

Current peak detection methods (such as the local maximum search plus threshold method) export detection results simply as peaks or non-peaks. Given the noisy nature of MS data, this simplification is prone to being influenced by noise (noise may also produce some local maximal values) and is very sensitive to specific parameter settings (including the intensity threshold value). In addition, a uniform threshold value may exclude some weak peaks in the MS data, while the existence/nonexistence of some weak peaks may be the most informative biomarkers.

Instead of using a binary output, it would be better to use both peak width and intensity information as quantitative measures of how likely it is that a candidate is a true peak. We can use a distribution model to describe the typical peak width and intensity. The parameters of the distribution can be estimated using training samples. Then a likelihood ratio test can be used to replace the binary peak detection result (either as peak (one) or as non-peak (zero)) with a real value. This new measure should provide richer information about peaks. We believe this will help us to better align multiple peak sets.

The challenge in peak alignment is that current methods may not work if we have large peak variation [like with LC/MS (liquid chromatography/mass spectrometry) data]. Another unsolved problem is that it may not be valid to assume that the distribution of peaks is not corrupted by noise (false positive detection).

To address these problems, we may consider the "true" locations of peaks as random variables and regard the peak detection results as sampling observations. Then, the problem of aligning multiple peak sets is converted to the problem of finding the mean (or median) values of random variables since we can assume that the majority of peaks should be located close to the "true" locations, with only a few outliers not obeying this assumption. After the mean/median values have been found/estimated, the remaining task is to simply align peaks w.r.t. the mean/median standard. Intuitively, the relative distance between a peak and its mean/median standard may also be used as a confidence measure in alignment.

34.3 Feature Selection

For current large-scale genomic and proteomic datasets, there are usually hundreds of thousands of features (also called variables in the following discussion) but limited sample size, which poses a unique challenge for statistical analysis. Feature selection serves two purposes in this context: biological interpretation and to reduce the impact of noise.

Suppose we have n_1 samples from one group (e.g. cancer patients) and n_0 samples from another group (e.g. normal subjects). We have m variables $(X_1, \ldots, X_m)$ (e.g. m/z ratios). For the kth variable, the observations are

$$X_k^1 = (x_{k1}, \ldots, x_{kn_1})$$

for the first group and

$$X_k^0 = \left(x_{k(n_1+1)}, \ldots, x_{k(n_1+n_0)}\right)$$

for the second group. They can be summarized in a data matrix, $X = (x_{ij})$. Assume X_k^1 are n_1 i.i.d. samples from one distribution $f_{k1}(x)$ and X_k^0 are n_0 i.i.d. samples from another distribution $f_{k0}(x)$.

Two sample t-test statistics or variants thereof are often used to quantify the difference between two groups in the analysis of gene expression data [34.18–20]

$$T_i = \frac{\bar{x}_{i1} - \bar{x}_{i0}}{\sqrt{\frac{1}{n_1}\hat{\sigma}_{i1}^2 + \frac{1}{n_0}\hat{\sigma}_{i0}^2}}, \tag{34.1}$$

where

$$\bar{x}_{i1} = \sum_{k=1}^{n_1} x_{ik}, \quad \bar{x}_{i0} = \sum_{k=n_1+1}^{n_1+n_0} x_{ik},$$

$$\hat{\sigma}_{i1}^2 = \frac{1}{n_1 - 1}\sum_{k=1}^{n_1}(x_{ik} - \bar{x}_{i1})^2,$$

$$\hat{\sigma}_{i0}^2 = \frac{1}{n_0 - 1}\sum_{k=n_1+1}^{n_1+n_0}(x_{ik} - \bar{x}_{i0})^2.$$

T_i can be interpreted as the standardized difference between these two groups. It is expected that the larger the standardized difference, the more separated the two groups are. One potential problem with using t-statistics is its lack of robustness, which may be a serious drawback when hundreds of thousands of features are being screened to identify informative ones.

34.3.1 A Simple Example of the Effect of Large Numbers of Features

Although there are hundreds of thousands of peaks representing peptides, we expect the number of peaks that provide information on disease classification to be limited. This, coupled with the limited number of samples available for analysis, poses great statistical challenges for the identification of informative peaks. Consider the following simple example: suppose that there are $n_1 = 10, n_2 = 10$, $m_0 = 10^3$ peptides showing no difference, and $m_1 = 40$ peptides showing differences between the two groups with $\lambda = \mu/\sigma = 1.0$. We can numerically calculate the expected number of significant features for these two groups

$$N_0 = 2m_0[1 - T(x, df = 18)] ,$$
$$N_1 = m_1[T(-x, df = 18, \lambda = 1.0) + 1 - T(x, df = 18, \lambda = 1.0)] ,$$

where $T(x, df)$ is the t-distribution function with df degrees of freedom, $T(x, df, \lambda)$ is the t-distribution function with df degrees of freedom and noncentral parameter λ, and the significance cut-off values are chosen as $|T| > x$. Figure 34.5 gives a comparison of true and false positives for this example, where a diagonal line is also shown. We can clearly see the dominant effect of noise in this example.

This artificial example reveals the difficulty that extracting useful features among a large number of noisy features entails. In practice, due to the noisy nature of MS data, the variance σ for individual peptide intensity will be very large, reflecting the difficulties with reproducibility that are commonly observed for MS data. Also, the number of noisy features m_0 (mostly uninformative) are increasing exponentially with the advance of technology (e.g. MALDI-FTICR data). The combination of these two factors will increase the ratio of false/true positives.

In this simple example, we ignore the interaction of different proteins. For complex diseases, such as cancer, it is quite possible that the effects result from the joint synergy of multiple proteins, while they individually show nonsignificant differences. Novel statistical methods are needed to account for the effects of noise and interactions among features.

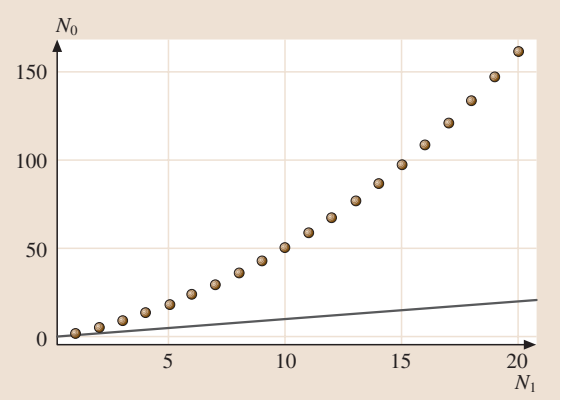

Fig. 34.5 Comparison of true positive and false positive for the simulation example. N_1 is the number of true positive, N_0 false positive. The diagonal line is also plotted as a *solid line*. Different points correspond to different settings of critical values in the t-test

34.3.2 Interaction

In ordinary or logistic regression models, we describe the interaction of different variables by including the interaction terms. This approach quickly becomes unfeasible with an increasing number of variables. Therefore, standard regression models are not appropriate due to $n \ll p$.

Instead of using univariate feature selection methods, it may be useful to consider multivariate feature selection methods. *Lai* et al. [34.21] analyzed the coexpression pattern of different genes using a prostate cancer microarray dataset, where the goal is to select genes that have differential gene–gene coexpression patterns with a target gene. Some interesting genes have been found to be significant and reported to be associated with prostate cancer, yet none of them showed marginal significant differential gene expressions.

Generally, multivariate feature selection is a combinatorial approach. To analyze two genes at a time we need to consider n^2 possibilities instead of n for the univariate feature selection. To analyze the interaction of K genes we need to consider n^K possibilities, which quickly becomes intractable.

A classification and regression tree (CART) [34.22] naturally models the interaction among different variables, and it has been successfully applied to genomic and proteomic datasets where $n \ll p$ is expected [34.23].

There are several new developments that are generalizations of the tree model. Bagging stands for bootstrap aggregating. Intuitively, bagging uses bootstrap to produce pseudoreplicates to improve prediction

accuracy [34.24]. The boosting method [34.25] is a sequential voting method, which adaptively resamples the original data so that the weights are increased for those most frequently misclassified samples. A boosting model using a tree as the underlying classifier has been successfully applied to genomic and proteomic datasets [34.26, 27].

34.3.3 Reducing the Influence of Noise

For most statistical models, the large number of variables may cause an overfitting problem. Just by chance, we may find some combinations of noise that can potentially discriminate samples with different disease status. We can incorporate some additional information into our analysis. For MS data, for instance, we only want to focus on peaks resulting from peptide/protein ionization. In previous sections we have addressed and emphasized the importance of MS data preprocessing.

34.3.4 Feature Selection with Machine Learning Methods

Isabelle et al. [34.28] have reported using SVM to select genes for cancer classification from microarray data. *Qu* et al. [34.29] applied a boosting tree algorithm to classify prostate cancer samples and to select important peptides using MS analysis of sera. *Wu* et al. [34.5] reported using random forest to select important biomarkers from ovarian cancer data based on MALDI-MS analysis of patient sera.

One distinct property of these learning-based feature selection methods compared to traditional statistical methods is the coupling of feature selection and sample classification. They implicitly approach the feature selection problem from a multivariate perspective. The significance of a feature depends strongly upon other features. In contrast, the feature selection methods employed in *Dudoit* et al. [34.30] and *Golub* et al. [34.31] are univariate and interactions among genes are ignored.

34.4 Sample Classification

There are many well established discriminant methods, including linear and quadratic discriminant analysis, and k-nearest neighbor, which have been compared in the context of classifying samples using microarray and MS data [34.5, 30]. The majority of these methods were developed in the pregenome era, where the sample size n was usually very large while the number of features p was very small. Therefore, directly applying these methods to genomic and proteomic datasets does not work. Instead, feature selection methods are usually applied to select some "useful" features at first and then the selected features are used to carry out sample classification based on traditional discriminant methods. This two-step approach essentially divides the problem into two separate steps: feature selection and sample classification, unlike the recently developed machine learning methods where the two parts are combined together.

The previously mentioned bagging (*Breiman*) [34.24], boosting (*Freund* and *Schapire*) [34.25], random forest (*Breiman*) [34.32], and support vector machine (*Vapnik*) [34.33] approaches have all been successfully applied to high-dimension genomic and proteomic datasets.

Due to the lack of a genuine testing dataset, cross-validation (CV) has been widely used to estimate the error rate for the classification methods. Inappropriate use of CV may seriously underestimate the real classification error rate. *Ambroise* and *McLachlan* [34.34] discussed the appropriate use of CV to estimate classification error rate, and recommended the use of K-fold cross-validation, e.g. $K = 5$ or 10.

34.5 Random Forest: Joint Modelling of Feature Selection and Classification

Wu et al. [34.5] compared the performance of several classical discriminant methods and some recently developed machine learning methods for analyzing an ovarian cancer MS dataset. In this study, random forest was shown to have good performance in terms of feature selection and sample classification. Here we design an algorithm to get an unbiased estimation of the classification error using random forest and at the same time efficiently extract useful features.

Suppose the preprocessed MS dataset has n samples and p peptides. We use $\{X_k \in \mathbb{R}^p,\ k = 1, 2, \cdots, n\}$ to represent the intensity profile of the kth individ-

ual, and $\{Y_k,\ k = 1, 2, \cdots, n\}$ to code the sample status.

The general idea of random forest is to combine random feature selection and bootstrap resampling to improve sample classification. We can briefly summarize the general idea as the following algorithm.

General random forest algorithm

1. Specify the number of bootstrap samples B, say 10^5.
2. For $b = 1, 2, \cdots, B$,
   a) Sample with replacement n samples from $\{X_k\}$ and denote the bootstrap samples by $\boldsymbol{X}^b = \{X_{b_1}, \cdots, X_{b_n}\}$, the corresponding response being $\boldsymbol{Y}^b = \{Y_{b_1}, \cdots, Y_{b_n}\}$.
   b) Randomly select m out of p peptides. Denote the selected subset of features by $\{r_1, \cdots, r_m\}$, and the bootstrap samples restricted to this subset by $\boldsymbol{X}^b_m$. Build a tree classifier T_b using $\boldsymbol{Y}^b$ and $\boldsymbol{X}^b_m$. Predict those samples not in the bootstrap samples using T_b.
3. Average the prediction over bootstrap samples to produce the final classification.

For the random forest algorithm from *Breiman* [34.32], randomness is introduced at each node split. Specifically, at each node split, a fixed number of features is randomly selected from all of the features and the best split is chosen among these selected features. For the random subspace method developed by *Ho* [34.35], a fixed number of features is selected at first and is used for the same original data to produce a tree classifier. Thus, both models have the effect of randomly using a fixed subset of features to produce a classifier, but differ in the underlying tree-building method.

Figure 34.6 shows a simple comparison of the two methods. We selected 78 peptides from the ovarian cancer MS data reported by *Wu* et al. [34.5]. Then we apply the two algorithms to numerically evaluate their sample classification performance using the selected subset of features. We want to emphasize that the calculated classification error rate is not a true error rate because we have used the sample status to select 78 peptides first. Our purpose here is just to show a simple numerical comparison of these two methods.

Other important issues in the analysis of MS data include specification of the number of biomarkers and the sample size being incorporated into the experimental design. To estimate the classification error Err, as discussed in *Cortes* et al. [34.36], the inverse power law learning curve relationship between Err and sample size N,

$$\mathrm{Err}(N) = \beta_0 N^{-\alpha} + \beta_1 \,, \tag{34.2}$$

is approximately true for large sample size datasets (usually about tens of thousands of samples); β_1 is the asymptotic classification error and (α, β_0) are positive constants.

Current MS data usually have a relatively small sample size ($N \approx 10^2$) compared to the high-dimension feature space ($p \approx 10^5$). In this situation, it may not be appropriate to rely on the learning curve model to extrapolate β_1, which corresponds to an infinite training sample size $N = \infty$. But within a limited range, this model may be useful to extrapolate the classification error. To estimate parameters $(\alpha, \beta_0, \beta_1)$, we need to obtain at least three observations.

Obviously the classification error Err also depends on the selected number of biomarkers m. We are going to use the inverse-power-law (34.2) to model $\mathrm{Err}(N, m)$.

We proposed the following algorithm to get an unbiased estimate for the classification error rate, which also provides an empirical method to select the number of biomarkers (Wu et al.) [34.37].

CV error estimation using random forest algorithm

1. Specify the number of folders K, say 5, and the range for the number of biomarkers m, say $M = \{20, 21, \cdots, 100\}$. Randomly divide all N samples into balanced K groups.
2. For $k = 1, 2, \cdots, K$ do the following:
   a) Use samples in the kth group as the testing set *Ts* and all the other samples as the training set *Tr*.

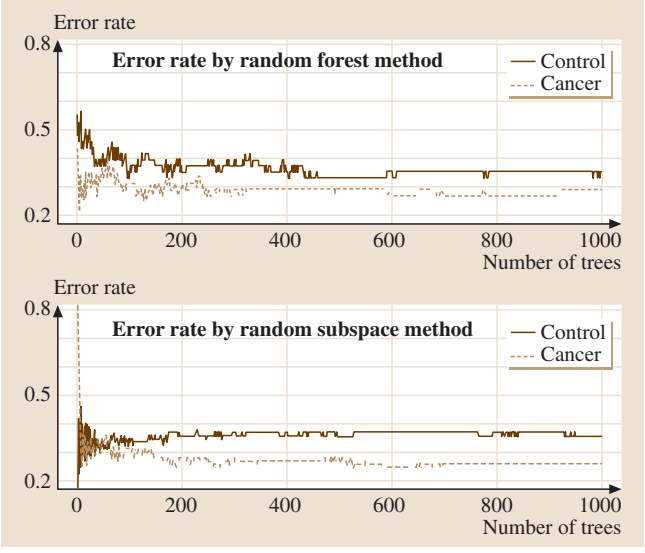

Fig. 34.6 Comparison of the error rates of two random forest algorithms on an ovarian cancer data set. 78 features selected by t-test are used in both algorithms. The two methods give similar performances

b) Apply the random forest algorithm (or any other feature selection method) to the training data Tr. Rank all of the features according to their importance.

c) Use the first $m \in M$ most important features and construct a classifier based on the training set Tr and predict samples in the testing set Ts. We will get a series of error estimates

$$\left\{\epsilon\left(k, m, \frac{K-1}{K}N\right), m \in M\right\} ,$$

where $\frac{K-1}{K}N$ is the effective size of the training set.

d) Use samples in the ith and jth group as the testing set and other $K-2$ groups as the training set. Repeat steps (2.2) and (2.3) to get the error estimate

$$\left\{\epsilon\left(k, m, \frac{K-2}{K}N\right), m \in M\right\} ,$$

where $\frac{K-2}{K}N$ is the effective size of the training set.

e) We can repeat step (2.4) using n of the groups as a testing set and get the error rate

$$\left\{\epsilon\left(k, m, \frac{K-n}{K}N\right), m \in M\right\} ,$$
$$n = 1, 2, \cdots, K-1 .$$

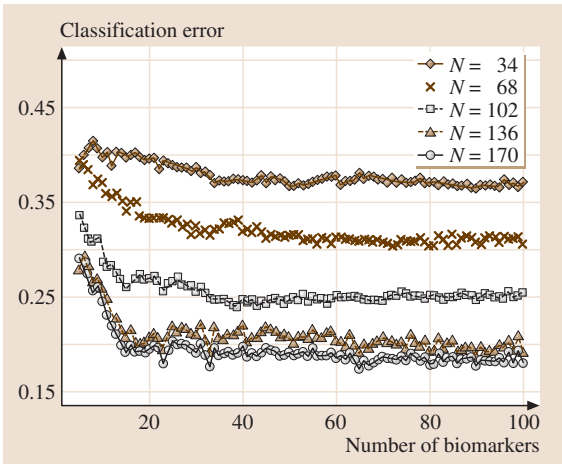

Fig. 34.7 Fivefold cross-validation estimation of classification error rate achieved by applying a random forest algorithm to the ovarian dataset. The error rates for sample size $N = 34, 68, 102, 136$ are obtained from the fivefold CV and the error rate for $N = 170$ is extrapolated from the inverse power law model fitting

3. Average $\epsilon[k, m, N(K-n)/K]$ over K folders to get the final error estimation $\bar{\epsilon}[m, N(K-n)/K]$ for m biomarkers and sample size $N(K-n)/K$.
4. Fit the inverse power law model (34.2) to $\bar{\epsilon}[m, N(K-n)/K]$ for every fixed m and extrapolate the error rate to N samples, $\bar{\epsilon}(m, N)$.

The estimated error rate curve $\bar{\epsilon}(m, n)$ can be used as a guidance for sample size calculation and to select the number of biomarkers.

For K folders, the previous algorithm will involve a total of 2^K training set fittings. If K is relatively large, say 10, the total number of fittings will be very large $(2^K = 1024)$. Note that in the inverse power law model (34.2) we only have three parameters $(\alpha, \beta_0, \beta_1)$. We can carry out just enough training data fitting, say 10, to estimate these three parameters. Then we can use the fitted model to interpolate or extrapolate the classification error rate for other sample size.

Figure 34.7 displays the fivefold CV estimate of the classification error rate achieved by applying the random forest algorithm to the serum mass spectrometry dataset for 170 ovarian cancer patients reported in *Wu* et al. [34.5], where the error rates for the training sample size $N = 34, 68, 102, 136$ are derived from the fivefold CV, and the error rate for $N = 170$ is extrapolated from the inverse power law model fitting.

34.5.1 Remaining Problems in Feature Selection and Sample Classification

As we discussed before, the univariate feature selection based on t-statistics is very sensitive to noise. We can reduce the influence of noise marginally by using additional biological information. But more importantly, we need to develop robust statistical methods. It is conjectured that random forest (*Breiman*) [34.32] does not over-fit. Our experience shows that we can dramatically reduce the classification error rate by incorporating feature selection with the random forest algorithm.

Intuitively, the sample classification error rate will increase with too much noise in the data. In this sense, feature selection will help us to improve the performance of algorithm classification. However, feature selection is usually affected by the small sample size $(n \ll p)$ in genomic and proteomic datasets. If we only select a small number of features, we may miss many "useful" features. One approach would be to couple the fast univariate feature selection with computationally intensive machine learning methods. For example, we can first use univariate feature selection to reduce the number of

features to a manageable size M_0. Then, we can apply the machine learning methods to refine our selection to a small number of target features M_1. Certainly, determining M_0 is a trade-off issue: if M_0 is too small, we will miss many informative features; if M_0 is too large, we will have a heavy computing burden for the following machine learning methods and also make the feature selection unstable.

For the genomic and proteomic data, the large n and small p will make the majority of the traditional statistical methods unusable. Most recently developed machine learning methods are computationally intensive and are often evaluated by empirical performance on some datasets. Statistical methods needed to be developed to bridge the traditional model-based principles and the newly developed machine learning methods.

34.6 Protein/Peptide Identification

34.6.1 Database Searching

MS in combination with database searching has emerged as a key platform in proteomics for high-throughput identification and quantification of proteins. A single stage MS provides a "mass fingerprint" of the peptide products of the enzymatically digested proteins, and this can be used to identify proteins [34.38–45]. This approach is useful for identifying purified proteins (such as proteins in dye-stained spots from 2-D polyacrylamide gels), and may also succeed with simple mixtures containing only 2–3 major proteins. Protein identification based on peptide mass database searching requires both high mass accuracy and that observed peptide masses be matched to a sufficient fraction (e.g. >25%) of the putatively identified protein. The latter task will be made more difficult if the protein has been post-translationally modified at multiple sites. Alternatively, the resulting peptide ions from the first stage MS can be isolated in the mass spectrometer and individually fragmented through CID to produce a tandem MS. In addition to the parent peptide mass, tandem MS provides structural information that can be used to deduce the amino acid sequences of individual peptides. Since tandem MS often identifies proteins using the CID-induced spectrum obtained from a single peptide, this technology is capable of identifying proteins in very complex mixtures such as cell extracts [34.43, 44, 46–61]. In general, database searching methods compare the experimentally observed tandem MS with features predicted for hypothetical spectra from candidate peptides (of equal mass) in the database and then return a ranked listing of the best matches, assuming that the query peptide exists in the protein sequence database. The statistical challenge in MS- and MS/MS-based protein identification is to assess the probability that a putative protein identification is indeed correct. In the case of MS/MS-based approaches, a commonly used criterion is that the observed MS/MS spectra must be matched to at least two different peptides from each identified protein.

34.6.2 De Novo Sequencing

An alternative approach to database searching of uninterpreted tandem MS for peptide identification is De Novo MS/MS sequencing [34.62–65], which attempts to derive a peptide sequence directly from tandem MS data. Although de novo MS/MS sequencing can handle situations where a target sequence is not in the protein database searched, the utility of this approach is highly dependent upon the quality of tandem MS data, such as the number of predicted fragment ion peaks that are observed and the level of noise, as well as the high level of expertise of the mass spectroscopist in interpreting the data, as there is no currently accepted algorithm capable of interpreting MS/MS spectra in terms of a de novo peptide sequence without human intervention. Because of the availability of DNA sequence databases, many of which are genome-level, and the very large numbers of MS/MS spectra (e.g., tens of thousands) generated in a single isotope coded affinity tag or another MS-based protein profiling analysis of a control versus experimental cell extract, highly automated database searching of uninterpreted MS/MS spectra is by necessity the current method of choice for high-throughput protein identification [34.43, 46, 49].

34.6.3 Statistical and Computational Methods

Due to the large number of available methods/alogrithms for MS- and MS/MS-based protein identification, we focus on what we believe are currently the most widely used approaches in the field.

- SEQUEST (Eng et al.) [34.46]

SEQUEST is one of the foremost yet sophisticated algorithms developed for identifying proteins from tandem

MS data. The analysis strategy can be divided into four major steps: data reduction, search for potential peptide matches, scoring peptide candidates and cross-correlation validation. More specifically, it begins with computer reduction of the experimental tandem MS data and only retains the most abundant ions to eliminate noise and to increase computational speed. It then chooses a protein database to search for all possible contiguous sequences of amino acids that match the mass of the peptide with a predetermined mass tolerance. Limited structure modifications may be taken into account as well as the specificity of the proteolytic enzyme used to generate the peptides. After that, SEQUEST compares the predicted fragment ions from each of the peptide sequences retrieved from the database with the observed fragment ions and assigns a score to the retrieved peptide using several criteria such as the number of matching ions and their corresponding intensities, some immonium ions, and the total number of predicted sequence ions. The top 500 best fit sequences are then subjected to a correlation-based analysis to generate a final score and ranking for the sequences.

SEQUEST correlates MS spectra predicted for peptide sequences in a protein database with an observed MS/MS spectrum. The cross-correlation score function provides a measure of the similarity between the predicted and observed fragment ions and a ranked order of relative closeness of fit of predicted fragment ions to other isobaric sequences in the database. However, since the cross-correlation score does not have probabilistic significance, it is not possible to determine the probability that the top-ranked and/or other matches result from random events and are thus false positives. Although lacking a statistical basis, *Eng* et al. [34.46] suggest that a difference greater than 0.1 between the normalized cross-correlation functions of the first- and second-ranked peptides indicates a successful match between the top-ranked peptide sequence and the observed spectrum. A commonly used guideline for Sequest-based protein identification is that observed MS/MS spectra are matched to two or more predicted peptides from the same protein and that each matched peptide meets the 0.1 difference criterion.

- MASCOT (*Perkins* et al.) [34.43]

MASCOT is another commonly used database searching algorithm, which incorporates a probability-based scoring scheme. The basic approach is to calculate the probability (via an approach that is not well described in the literature) that a match between the experimental MS/MS data set and each sequence database entry is a chance event. The match with the lowest probability of resulting from a chance event is reported as the best match. MASCOT considers many key factors, such as the number of missed cleavages, both quantitative and nonquantitative modifications (the number of nonquantitative modifications is limited to four), mass accuracy, the particular ion series to be searched, and peak intensities. Hence, MASCOT iteratively searches for the set with the most intense ion peaks, which provide the highest score – with the latter being reported as $-10\log(P)$, where P is the probability of the match resulting from a random, chance event. *Perkins* et al. [34.43] suggested that the validity of MASCOT probabilities should be tested by repeating the search against a randomized sequence database and by comparing the MASCOT results with those obtained via the use of other search engines.

- Other Methods

In addition to SEQUEST and MASCOT, many other methods have been proposed to identify peptides and proteins from tandem MS data. They range from the development of probability-based scoring schemes, the identification of modified peptides, and checking the identities of peptides and proteins in other miscellaneous fields. Here we give a brief review of these approaches.

Bafna and *Edwards* [34.49] proposed the use of SCOPE to score a peptide with a conditional probability of generating the observed spectrum. SCOPE models the process of tandem MS spectrum generation using a two-step stochastic process. Then SCOPE searches a database for the peptide that maximizes the conditional probability. The SCOPE algorithm works only as well as the probabilities assumed for each predicted fragment of a peptide. Although *Bafna* and *Edwards* [34.49] proposed using a human-curated database of identified spectra to compute empirical estimates of the fragmentation probabilities required by this algorithm, to our knowledge this task has not yet been carried out. Thus, SCOPE is not yet a viable option for most laboratories.

Pevzner et al. [34.48] implemented spectral convolution and spectral alignment approaches to identifying modified peptides without the need for exhaustive generation of all possible mutations and modifications. The advantages of these approaches come with a tradeoff in the accuracy of their scoring functions, and they usually serve as filters to identify a set of "top-hit" peptides for further analysis. *Lu* and *Chen* [34.60] developed a suffix tree approach to reduce search time when identifying modified peptides, but the resulting scores do not have direct probabilistic interpretations.

PeptideProphet [34.53] and ProteinProphet (*Nesvizhskii* et al.) [34.61] were developed at the Institute for

Systems Biology (ISB) to validate peptide and protein identifications using robust statistical models. After scores are derived from the database search, PeptideProphet models the distributions of these scores as a mixture of two distributions, with one consisting of correct matches, and the other consisting of incorrect matches. ProteinProphet takes as input the list of peptides along with probabilities from PeptideProphet, adjusts the probabilities for observed protein grouping information, and then discriminates correct from incorrect protein identifications.

Mann and *Wilm* [34.47] proposed a "peptide sequence tag" approach to extracting a short, unambiguous amino acid sequence from the peak pattern that, when combined with the mass information, infers the composition of the peptide. *Clauser* et al. [34.44] considered the impact of measurement accuracy in protein identification. *Kapp* et al. [34.66] proposed two different statistical methods, the cleavage intensity ratio (CIR) and a linear model, to identify the key factors that influence peptide fragmentation. It has been known for a long time that peptides do not fragment equally and that some bonds are more likely to break than others. However, the chemical mechanisms and physical processes that govern the fragmentation of peptides are highly complex. One can only take results from previous experiments and try to find indicators about such mechanisms. The use of these statistical methods demonstrates that proton mobility is the most important factor. Other important factors include local secondary structure and conformation as well as the position of a residue within a peptide.

34.7 Conclusion and Perspective

While the algorithms for protein identification from tandem MS mentioned above have different emphases, they contain the elements of the following three modules [34.49]:

1. Interpretation [34.67], where the input MS/MS data are interpreted and the output may include parent peptide mass and possibly a partial sequence.
2. Filtering, where the interpreted MS/MS data are used as templates in a database search in order to identify a set of candidate peptides.
3. Scoring, where the candidate peptides are ranked with a score.

Among these three modules, a good scoring scheme is the mainstay. Most database searching algorithms assign a score function by correlating the uninterpreted tandem MS with theoretical/simulated tandem MS for certain peptides derived from protein sequence databases. An emerging issue is the significance of the match between a peptide sequence and tandem MS data. This is especially important in multidimensional LC/MS-based protein profiling where, for instance, our isotope-coded affinity tag studies on crude cell extracts typically identify and quantify two or more peptides from only a few hundred proteins as compared to identifying only a single peptide from a thousand or more proteins. Currently, we require that two or more peptides must be matched to each identified protein. However, if statistically sound criteria could be developed to permit firm protein identifications based on only a single MS/MS spectrum, the useable data would increase significantly. Therefore, it is important and necessary to develop the best possible probability-based scoring schemes, particularly in the case of the automated high-throughput protein analyses used today.

Even for the probability-based algorithms, the efficiencies of score functions can be further improved by incorporating other important factors. For example, statistical models proposed by *Kapp* et al. [34.66] may be used to predict the important factors that govern the fragmentation pattern of peptides and subsequently improve the fragmentation probability as well as the score function in SCOPE [34.49]. In addition, some intensity information can be added to improve score function.

One common drawback of all of these algorithms is the lack of ability to detect modified peptides. Most of the database search methods are not mutation- and modification-tolerant. They are not effective at detecting types and sites of sequence variations, leading to low score functions. A few methods have incorporated mutation and modification into their algorithms, but they can only handle at most two or three possible modifications. Therefore, the identification of modified peptides remains a challenging problem. Theoretically, one can generate a virtual database of all modified peptides for a small set of modifications and match the spectrum against this virtual database. But the size of this virtual database increases exponentially with the number of modifications allowed, making this approach unfeasible. Markov chain Monte Carlo is an appealing approach to identifying mutated and modified peptides. The algorithm may start from a peptide corresponding

to a protein and a "new" candidate peptide with modifications/mutations is proposed according to a set of prior probabilities for different modifications and mutations. The proposed "new" peptide is either rejected or accepted and the procedure can be iterated to sample the posterior distribution for protein modification sites and mutations. However, the computational demands can also be enormous for this approach. Parallel computation and better-constructed databases are necessary to make this approach more feasible.

References

34.1 D. Greenbaum, C. Colangelo, K. Williams, M. Gerstein: Computing protein abundance and mRNA expression levels on a genomic scale, Genome Biol. **4**, 117.1–117.8 (2003)

34.2 M. Wagner, D. Naik, A. Pothen: Protocols for disease classification from mass spectrometry data, Proteomics **3**(9), 1692–1698 (2003)

34.3 Y. Yasui, M. Pepe, M.L. Thompson, B. Adam, G.L. Wright Jr., Y. Qu, J.D. Potter, M. Winget, M. Thornquist, Z. Feng: A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection, Biostatistics **4**(3), 449–463 (2003)

34.4 K.R. Coombes, H.A. Fritsche, Jr, C. Clarke, J. Chen, K.A. Baggerly, J.S. Morris, L. Xiao, M. Hung, H.M. Kuerer: Quality control, peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption, ionization, Clinical Chemistry **49**(10), 1615–1623 (2003)

34.5 B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, H. Zhao: Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data, Bioinformatics **19**(13), 1636–1643 (2003)

34.6 Q. Liu, B. Krashnapuram, P. Pratapa, X. Liao, A. Hartemink, L. Carin: Identification of differentially expressed proteins using maldi-tof mass spectra. In: *ASILOMAR Conference: Biological Aspects of Signal Processing* 2003)

34.7 Y. Yasui, D. McLerran, B.L. Adam, M. Winget, M. Thornquist, Z.D.Z.D. Feng: An automated peak identification/calibration procedure for high-dimensional protein measures from mass spectrometers, J. Biomed. Biotec. **4**, 242–248 (2003)

34.8 G.A. Satten, S. Datta, H. Moura, A.R. Woolfitt, G. Carvalho, R. Facklam, J.R. Barr: Standardization and denoising algorithms for mass spectra to classify whole-organism bacterial specimens, Bioinformatics **20**(17), 3128–3136 (2004)

34.9 K.R. Coombes, S. Tsavachidis, J.S. Morris, K.A. Baggerly, M. Hung, H.M. Kuerer: *Improved peak detection, quantification of mass spectrometry data acquired from surface-enhanced laser desorption, ionization by denoising spectra with the undecimated discrete wavelet transform*, Technical report (Univ. Texas M.D. Anderson Cancer Center, Houston 2004)

34.10 T.W. Randolph and Y. Yasui: Multiscale processing of mass spectrometry data, University of Washington Biostatistics Working Paper Series, Number 230, (2004)

34.11 W. Yu, B. Wu, N. Lin, K. Stone, K. Williams, H. Zhao: Detecting, aligning peaks in mass spectrometry data with applications to MALDI, Comput. Biol. Chem. (2005) in press

34.12 R.J.O. Torgrip, M. Aberg, B. Karlberg, S.P. Jacobsson: Peak alignment using reduced set mapping, J. Chemometrics **17**, 573–582 (2003)

34.13 P.H.C. Eilers: Parametric time warping, Analytical Chemistry **76**(2), 404–411 (2004)

34.14 R. Tibshirani, T. Hastie, B. Narasimhan, S. Soltys, G. Shi, A. Koong, Q. Le: Sample classification from protein mass spectrometry, by "peak probability contrasts", Bioinformatics **20**(17), 3034–3044 (2004)

34.15 K.J. Johnson, B.W. Wright, K.H. Jarman, R.E. Synovec: High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis, J. Chromatography A **996**, 141–155 (2003)

34.16 N.V. Nielsen, J.M. Carstensen, J. Smedsgaard: Aligning of single, multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping, J. Chromatography A **805**, 17–35 (1998)

34.17 J. Aach, G.M. Church: Aligning gene expression time series with time warping algorithms, Bioinformatics **17**(6), 495–508 (2001)

34.18 S. Dudoit, Y.H. Yang, T.P. Speed, M.J. Callow: Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, Stat. Sinica **12**(1), 111–139 (2002)

34.19 V.G. Tusher, R. Tibshirani, G. Chu: Significance analysis of microarrays applied to the ionizing radiation response, Proc. Natl. Acad. Sci. **98**(9), 5116–5121 (2001)

34.20 X. Cui, G.A. Churchill: Statistical tests for differential expression in cDNA microarray experiments, Genome Biology **4**(4), 210 (2003)

34.21 Y. Lai, B. Wu, L. Chen, H. Zhao: Statistical method for identifying differential gene–gene coexpression patterns, Bioinformatics **20**(17), 3146–3155 (2004)

34.22 L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone: *Classification and Regression Trees* (Kluwer Academic, 1984)

34.23 E. C. Gunther, D. J. Stone, R. W. Gerwien, P. Bento, M. P. Heyes: Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro, Proc. Natl. Acad. Sci **100**(16), 9608–9613 (2003)

34.24 L. Breiman: Bagging predictors, Machine Learning **24**, 123–140 (1996)

34.25 Y. Freund, R. Schapire: A decision-theoretic generalization of online learning, an application to boosting, J. Computer, System Sci. **55**(1), 119–139 (1997)

34.26 B. Adam, Y. Qu, J. W. Davis, M. D. Ward, M. A. Clements, L. H. Cazares, O. J. Semmes, P. F. Schellhammer, Y. Yasui, Z. Feng: Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men, Cancer Research **62**(13), 3609–3614 (2002)

34.27 M. Dettling, P. Buhlmann: Boosting for tumor classification with gene expression data, Bioinformatics **19**(9), 1061–1069 (2003)

34.28 G. Isabelle, W. Jason, B. Stephen, V. Vladimir: Gene selection for cancer classification using support vector machines, Machine Learning **46**(1-3), 389–422 (2002)

34.29 Y. Qu, B. L. Adam, Y. Yasui, M. D. Ward, L. H. Cazares, P. F. Schellhammer, Z. Feng, O. J. Semmes, G. L. Wright Jr.: Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients, Clin. Chem. **48**(10), 1835–1843 (2002)

34.30 S. Dudoit, J. Fridlyand, T. P. Speed: Comparison of discrimination methods for the classification of tumors using gene expression data, J. Am. Stat. Assoc. **97**(457), 77–87 (2002)

34.31 T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, E. S. Lander: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science **286**(5439), 531–537 (1999)

34.32 L. Breiman: Random forests, Machine Learning **45**(1), 5–32 (2001)

34.33 V. N. Vapnik: *Statistical Learning Theory* (Wiley-Interscience, New York 1998)

34.34 C. Ambroise, G. J. McLachlan: Selection bias in gene extraction on the basis of microarray gene-expression data, Proc. Natl. Acad. Sci. **99**(10), 6562–6566 (2002)

34.35 T. K. Ho: The random subspace method for constructing decision forests, IEEE Trans. Pattern Anal. Mach. Intell. **20**(8), 832–844 (1998)

34.36 C. Cortes, L. D. Jackel, S. A. Solla, V. Vapnik, J. S. Denker: Learning curves: asymptotic values, rate of convergence, Adv. Neural Info. Proc. Systems **6**, 327–334 (1994)

34.37 B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, H. Zhao: Ovarian cancer classification based on mass spectrometry analysis of sera, Cancer Informatics (2005) in press

34.38 W. J. Henzel, T. M. Billeci, J. T. Stults, S. C. Wong, C. Grimley, C. Watanabe: Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases, Proc. Natl. Acad. Sci. **90**, 5011–5015 (1993)

34.39 P. James, M. Quadroni, E. Carafoli, G. Gonnet: Protein identification by mass profile fingerprinting, Biochem. Biophys. Res. Commun. **195**, 58–64 (1993)

34.40 M. Mann, P. Hojrup, P. Roepstorff: Use of mass spectrometric molecular weight information to identify proteins in sequence databases, Biol. Mass Spectrom. **22**, 338–345 (1993)

34.41 D. J. Pappin, P. Hojrup, A. J. Bleasby: Rapid identification of proteins by peptide-mass fingerprinting, Curr. Biol. **3**, 327–332 (1993)

34.42 J. R. Yates III, S. Speicher, P. R. Griffin, T. Hunkapiller: Peptide mass maps: A highly informative approach to protein identification, Anal. Biochem. **214**, 397–408 (1993)

34.43 D. N. Perkins, D. J. Pappin, D. M. Creasy, J. S. Cottrell: Probability-based protein identification by searching sequence databases using mass spectrometry data, J. S. Electrophoresis **20**, 3551–3567 (1999)

34.44 K. R. Clauser, P. Baker, A. I. Burlingame: Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching, Anal. Chem. **71**, 2871–2882 (1999)

34.45 W. Zhang, B. T. Chait: ProFound: An expert system for protein identification using mass spectrometric peptide mapping information, Anal. Chem. **72**, 2482–2489 (2000)

34.46 J. K. Eng, A. L. McCormack, J. R. Yates: An approach to correlate MS/MS data to amino acid sequences in a protein database, J. Am. Soc. Mass Spectrom. **5**, 976–989 (1994)

34.47 M. Mann, M. S. Wilm: Error-tolerant identification of peptides in sequence databases by peptide sequence tags, Anal. Chem. **66**, 4390–4399 (1994)

34.48 P. A. Pevzner, V. Dancik, C. L. Tang: Mutation-tolerant protein identification by mass spectrometry, J. Comput. Biol. **7**, 777–787 (2000)

34.49 V. Bafna, N. Edwards: SCOPE: A probabilistic model for scoring tandem mass spectra against a peptide database, Bioinformatics **17**, S13–21 (2001)

34.50 B. T. Hansen, J. A. Jones, D. E. Mason, D. C. Liebler: SALSA: A pattern recognition algorithm to detect electrophile-adducted peptides by automated

evaluation of CID spectra in LC-MS-MS analyses, Anal. Chem. **73**, 1676–1683 (2001)

34.51 D. M. Creasy, J. S. Cottrell: Error-tolerant searching of uninterpreted tandem mass spectrometry data, Proteomics **2**, 1426–1434 (2002)

34.52 H. I. Field, D. Fenyo, R. C. Beavis: RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in arelational database, Proteomics **2**, 36–47 (2002)

34.53 A. Keller, A. I. Nesvizhskii, E. Kolker, R. Aebersold: Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search, Anal. Chem. **74**, 5389–5392 (2002)

34.54 M. J. MacCoss, C. C. Wu, J. R. Yates: Probability-based validation of protein identifications using amodified SEQUEST algorithm, Anal. Chem. **74**, 5593–5599 (2002)

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

34.56 J. Colinge, A. Masselot, M. Giron, T. Dessigny, J. Magnin: OLAV: towards high throughput tandem mass spectrometry data identification, Proteomics **3**, 1454–1463 (2003)

34.57 E. Gasteiger, A. Gattiker, C. Hoogland, I. Ivanyi, R. D. Appel, A. Bairoch: ExPASy: The proteomics server for in-depth protein knowledge and analysis, Nucleic Acids Res. **3**, 3784–3788 (2003)

34.58 M. Havilio, Y. Haddad, Z. Smilansky: Intensity-based statistical scorer for tandem mass spectrometry, Anal. Chem. **75**, 435–444 (2003)

34.59 P. Hernandez, R. Gras, J. Frey, R. D. Appel: Popitam: towards new heuristic strategies to improve protein identification from tandem mass spectrometry data, Proteomics **3**, 870–878 (2003)

34.60 B. Lu, T. Chen: A suffix tree approach to the interpretation of tandem mass spectra: applications to peptides of non-specific digestion, post-translational modifications, Bioinformatics **19**, 113–121 (2003)

34.61 A. I. Nesvizhskii, A. Keller, E. Kolker, R. Aebersold: A statistical model for identifying proteins by tandem mass spectrometry, Anal. Chem. **75**, 4646–4658 (2003)

34.62 J. A. Taylor, R. S. Johnson: Sequence database searches via de novo peptide sequencing by tandem mass spectrometry, Rapid Commun. Mass Spectrom. **11**, 1067–75 (1997)

34.63 V. Dancik, T. A. Addona, K. R. Clauser, J. E. Vath, P. A. Pevzner: De Novo peptide sequencing via tandem mass spectrometry, J. Comput. Biol. **6**, 327–342 (1999)

34.64 T. Chen, M. Y. Kao, M. Tepel, J. Rush, G. M. Church: A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry, J. Comput. Biol. **8**, 325–337 (2001)

34.65 B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, G. Lajoie: PEAKS: Powerful software for peptide de novo sequencing by tandem mass spectrometry, Rapid Commun. Mass Spectrom. **17**, 2337–2342 (2003)

34.66 E. A. Kapp, F. Schütz, G. E. Reid, J. S. Eddes, R. L. Moritz, R. A. J. O'Hair, T. P. Speed, R. J. Simpson: Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation, Anal. Chem. **75**, 6251–6264 (2003)

34.67 D. C. Chamrad, G. Koerting, J. Gobom, H. Thiele, J. Klose, H. E. Meyer, M. Blueggel: Interpretation of mass spectrometry data for high-throughput proteomics, Anal. Bioanal. Chem. **376**, 1014–1022 (2003)