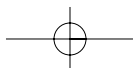
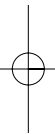
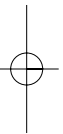
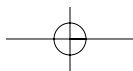
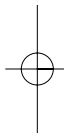
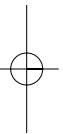
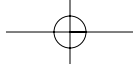


Section VII

New technologies





Serum Proteomic Profiling and Analysis

Chapter 57

**Richard Pelikan¹, Michael T. Lotze², Jim Lyons-Weiler¹,
David Malehorn¹, Milos Hauskrecht¹**

¹*Department of Computer Science, University of Pittsburgh;*

²*Translational Research Molecular Medical Institute,
University of Pittsburgh School of Medicine, Pittsburgh, PA, USA*

They killed him in Sarajevo, Mr Švejk. They shot him with a revolver as he was riding with that archduchess of his in an automobile.

Jaroslav Hašek *Fateful Adventures of the Good Soldier Švejk
During the World War
Osudy dobrého vojáka Švejka za světové války*

INTRODUCTION

The ability to examine serum peptides and proteins using mass spectrometry (MS) has recently become broadly of interest as a novel biomarker and surrogate of disease discovery tool. Just as Švejk's world was transformed by the introduction of modern technology now, almost a century later, the introduction of modern mass spectrometry strategies, assessing data-dense putative markers associated with inflammatory or immune endpoints, have indeed changed the world of biological investigation. They have become more widely applied, particularly in the setting of cancer diagnostics. Surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF MS) proteomic profiling is one of these increasingly popular tools, using 'shooting' of protein mixtures with focused lasers, in the search for known and surrogate biomarkers for disease diagnosis and prognosis. Current diagnostic tests, such as repetitive biopsies done before or after therapy, impose greater costs and risk of injury as opposed to gathering the patient's proteomic

profile. The greatest potential promise of proteomic profiling lies in the possibility of the early detection of a malignant or chronic inflammatory condition with simple tests of the serum or urine. In addition, the hope is that different disease types and response to therapy phenotypes might be reproducibly distinguishable using rapid and relatively inexpensive assays.

SELDI-TOF-MS rapidly provides a protein expression profile from a variety of biological and clinical samples. The potential efficacy of this system for serum protein profiling of cancer in human breast (Paweletz et al., 2001), colon (Watkins et al., 2001), head and neck (Wadsworth et al., 2004), lung (Zhukov et al., 2003), ovarian (Petricoin et al., 2002), prostate cancer (Adam et al., 2002; Petricoin and Ornstein, 2002) and hepatoma (Steel et al., 2003; Zeindl-Eberhart et al., 2004) has been recently demonstrated. These studies describe diagnostic features of these profiles and classification algorithms based on these features, which provide at least 80 per cent and, in some cases, >90 per cent classification accuracy between cancer cases and controls. Papers reporting high sensitivity and specificity in class prediction presented promising initial positive results. Efforts to relate the proteomic profiling to changes in known protein factors, including cytokines and chemokines, as well as cellular elements and their signaling capacity, are now being explored, but it is premature to include them in this chapter.

The goal of this chapter is to overview the proteome profiling technology and its potential benefits for identification and clinical assessment of progression of

pathological conditions. The initial discussion of the SELDI-TOF technology is followed by a discussion of technical limitations that affect the interpretive analysis of the profiles. Next we focus on the description of some statistical methods used to decipher the profiles and their usage in diagnosing or predicting the condition of patients. We illustrate the potential of these methods on the task of differentiating between individuals with and without cancer. Finally, the review concludes with some insight on the direction of using proteomic data analysis towards the benefit of the medical community.

SELDI-TOF MASS SPECTROMETRY

The ProteinChip® Biology System developed by Ciphergen Biosystems, Inc. uses SELDI-TOF MS to ionize proteins specifically retained on a chromatographic surface, which are then detected by time-of-flight mass spectrometry. The system can be used for the mass analysis of compounds such as proteins, peptides and nucleic acids within a range of 0–200 kD. The procedure begins with the reaction of a biological sample (e.g. bodily fluid, cell lysate or fraction thereof) with the chromatographic surface (or 'spot') of a ProteinChip, which possesses a defined affinity characteristic: anionic/cationic, hydrophobic, or metal-binding, or biologically derivatized (e.g. antibody coupled). The ProteinChips, comprised of 8 or 16 of these spots, retain only those analytes that match the surface's physical affinity characteristics – non-binding species are washed away using appropriate conditions. The spots are then overlaid with an energy-absorbing 'matrix' compound, which co-crystallizes around the retained analyte molecules. The spots are 'shot' multiple times by a pulsed nitrogen laser. The laser desorption

process results in ionization of matrix molecules and protonation of intact analyte molecules. The ions produced are differentially accelerated in an electrical field and then detected after passing through a field-free, evacuated 'drift' tube. The time of flight across the tube is converted to provide information on the molecule's mass-to-charge ratio (m/z), since heavier molecules, by Newtonian physics, will take longer to travel the same distance, having acquired less initial velocity from the uniform acceleration force. The detected ions are then represented as a 'spectrum' with peaks of varying intensities and molecular weight assignments are made relative to known calibrant species. Figure 57.1 displays a summary of the SELDI-TOF MS process.

Early studies and first applications (Paweletz et al., 2001; Petricoin and Ornstein, 2002; Petricoin et al., 2002) assembled SELDI-TOF MS proteomic profiles of patients with various types of cancer. The primary goal of such studies was to determine whether it is possible to detect peptide markers of the presence of disease by analyzing the profiles contrasting those with cancer and those without. For, example, Petricoin and coworkers' April 2002 study (Petricoin et al., 2002) compared profiles of 200 patients in order to determine a discriminating pattern between patients with ovarian cancer and those with a variety of non-cancer conditions. A sample profile from this set is shown in Figure 57.2. It consists of intensities measured over 15 154 mass/charge (m/z) values.

LIMITATIONS OF THE SELDI-TOF PROFILES

The profiles obtained by the SELDI-TOF MS system manifest a number of attributes which can complicate analysis. Figure 57.3 compares two unprocessed profiles from the

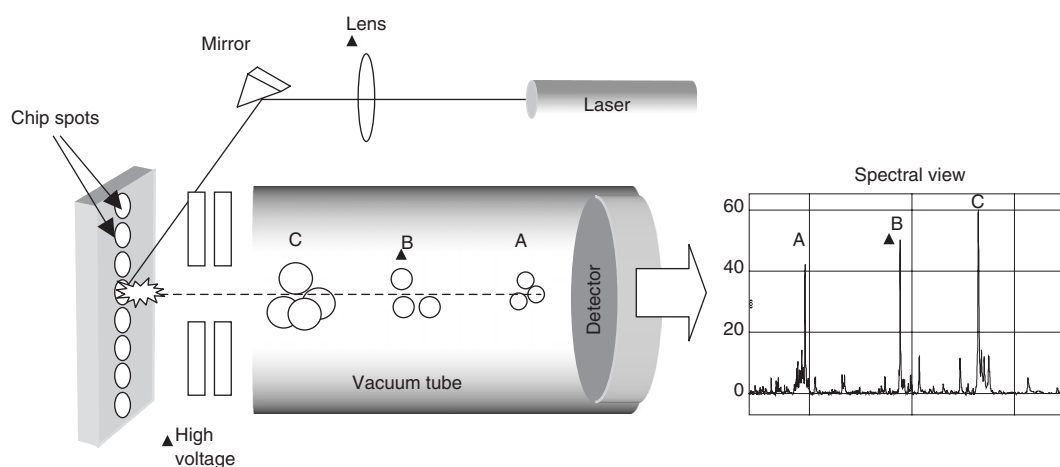


Figure 57.1 Diagram of the Ciphergen SELDI-TOF MS system. Samples are reacted with ProteinChip spots, coated with 'matrix' and pulsed with a laser. The ionized species created during this process are differentially accelerated in an electric field, float through the vacuum tube, where their arrival times and quantities (intensity) are measured by a detector. Heavier ions demonstrate longer time-of-flight, which is a unique indicator for each molecular species. The plot of ion intensity versus specific time of flight (or corresponding mass to charge value) constitutes the mass spectrum for a given sample.

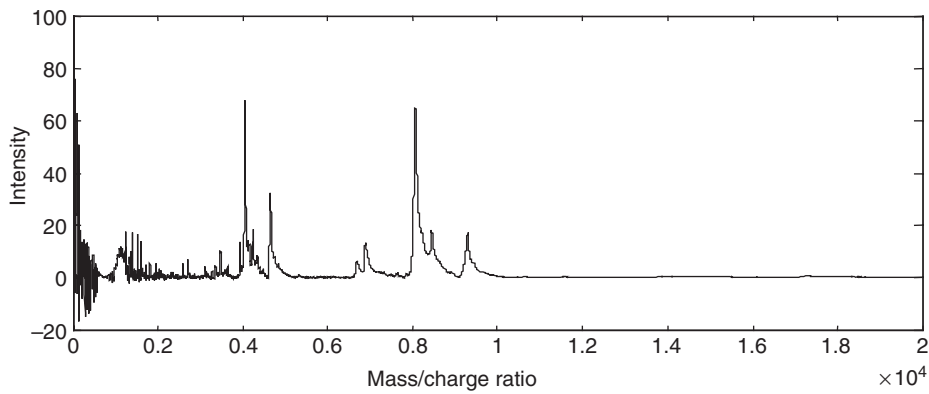


Figure 57.2 A sample SELDI-TOF MS profile. The x-axis plots mass-to-charge ratio. The y-axis plots the relative intensity (flux) of analyte species with the identified mass-to-charge ratio.

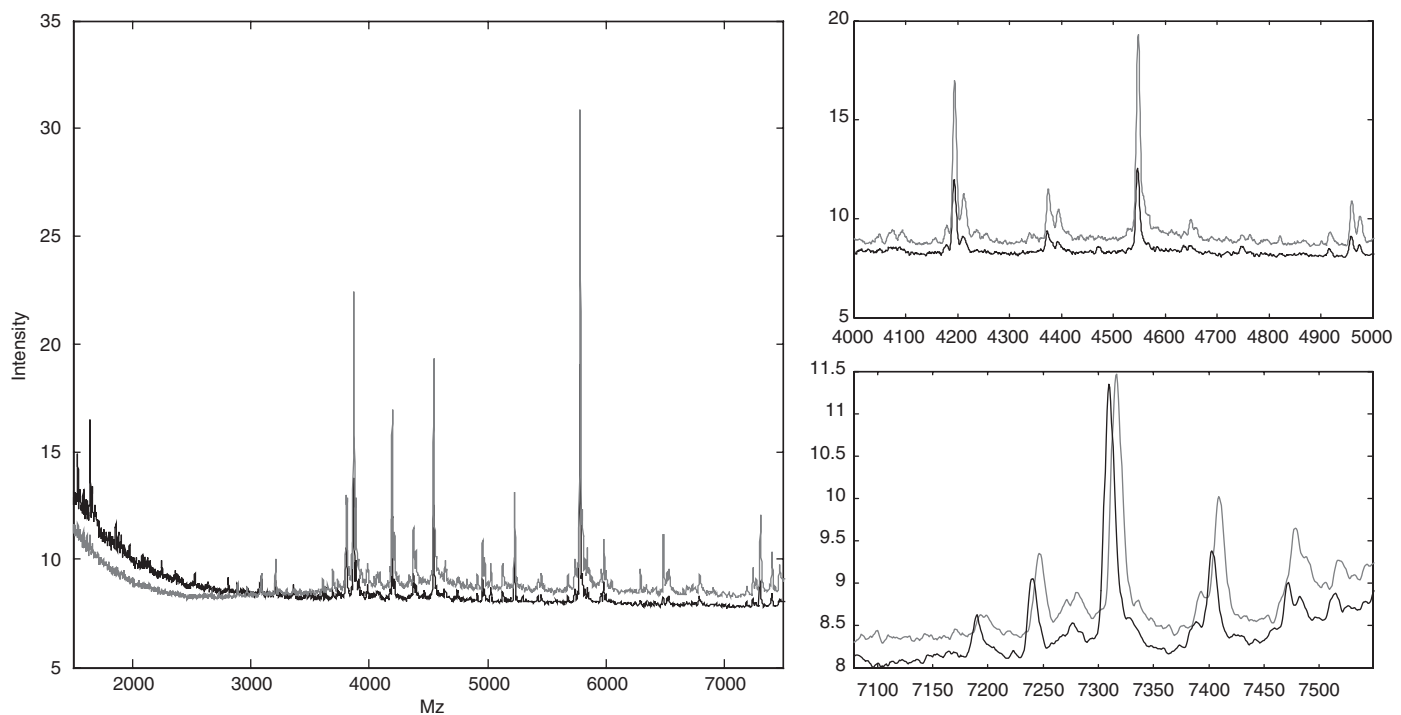


Figure 57.3 Two SELDI-TOF profiles obtained for the same pooled reference serum. The differences in the baseline (left panel), intensity measurements (right top panel) and mass inaccuracies (right bottom panel) are apparent.

same reference serum. Differences are readily visible and illustrate stochastic variations that obscure what would ideally be identical profiles. Many causes may be responsible for the differences: variation in the SELDI-TOF MS instrumentation condition over time, fluctuation in the intensity of the laser and even surface irregularities on the protein chip spots or the matrix crystallization.

All stochastic variations in profiles show up as differences in intensity readings. However, these differences are the result of two intertwined problems: *mass inaccuracy* and *intensity measurement errors*. Mass inaccuracy refers to the misalignment of readings for different m/z values. The mass inaccuracy for CIPHERGEN's SELDI-TOF MS system is reported to be approximately 0.1 per cent for externally calibrated experiments. The intensity measurement error may arise from imperfect performance of

the ion detector in registering the abundance of ions at a given time point (detector saturation). Both types of errors are illustrated in the right panel of Figure 57.3. In addition, the left panel of Figure 57.3 illustrates baseline variation, a systematic intensity measurement error for which the measurements of the profile differ from 0. Note that the baseline shifts between two samples differ despite the fact that the same serum is being analyzed.

The *mass inaccuracy* and *intensity measurement errors* can lead to significant fluctuation in profile readings. In addition, if we analyze samples from multiple individuals a natural biological variation in sera is observed. This can show up as differences in intensity values or as the presence or absence of peaks in the profile. The peaks are believed to indicate the presence of peptides or their fragments. These problems lead to serious challenges in

interpretive analysis. Many of the observed variations could be caused by changes in carrier proteins, protein catabolism or the inherent cyclical shifts in relative abundance due to production and release from various tissues. We have considered some of these issues, particularly during nominally chaotic states as represented in patients with various neoplasms reflecting on the notion that an individual analyte may be increased or decreased at any time interval when compared to normal sera, something we have termed the ABA problem (Patel and Lyons-Weiler, 2004).

DATA PREPROCESSING

Despite the problems presented above, there are possible steps one can take before analyzing the data. The cleansing and modifying, or *preprocessing*, of the data are intended to eliminate noise or redundant components in the signal other than true biological variation in the serum. The preprocessing steps include *smoothing*, *profile alignment*, *rescaling (normalization)* and *baseline correction*. However, any preprocessing step comes at a risk of loss of useful biological information or introduction of additional errors. Therefore, if any preprocessing is performed, conclusions drawn from an ensuing analysis should be carefully validated.

Smoothing serves to eliminate a high frequency noise component in the signal. Figure 57.4 illustrates the effect of smoothing on a SELDI-TOF profile. High frequency variation is eliminated by local averaging of the signal. Whether smoothing removes useful signal or useless noise is not yet clear.

The mass inaccuracy problem (see Figure 57.3, lower right panel) can be resolved through *profile alignment* methods. A number of strategies for performing profile alignment exist. One option is to define a reference

profile in terms of a set of established biomarkers that are easily identifiable in every profile. Another approach is to include indicator peptides in the serum, in order purposely to populate the profile with peaks to be expected at certain m/z values. The intensity readings between these peaks could then be stretched or shrunk along the x -axis appropriately. Unfortunately, due to the locality of m/z errors, this approach requires the addition of several thousand peptides to the serum in order to recapture properly the information lost through mass inaccuracy. Using alignment algorithms directly on the profiles runs the risk of eliminating important biological variability in the data. Clearly, this challenge is deserving of additional attention; solving the alignment problem entails the possibility of eliminating sample-to-sample variation, which appears to be autocorrelated in terms of temporal (run-to-run) and spatial (spot-to-spot) processing.

A particular profile may suffer from an overall weakness in signal. Variation in the sensitivity of the ion detector or amount of retained molecules on the chip surface may result in profiles which seem to be measured on a different scale (see Figure 57.3, upper right panel). *Rescaling (normalizing)* these profiles allows them to be compared on the same scale. This type of adjustment differs from baseline shifting, which is an additive error, as opposed to scaling, which is a multiplicative error.

Correcting for *baseline shifts* (see Figure 57.3, left panel) involves subtracting a constant intensity value from the profile. The problem is that the baseline shift may vary over the m/z range. Figure 57.5 illustrates the process of baseline correction on the reference profile. The method removed the additive component in the signal and brought the baseline to 0. A challenge presented by baseline correction is that the noise from the intensity measurement process appears to be correlated strongly with the magnitude of the measurement. This suggests that any signal rescaling must be performed before the baseline correction.

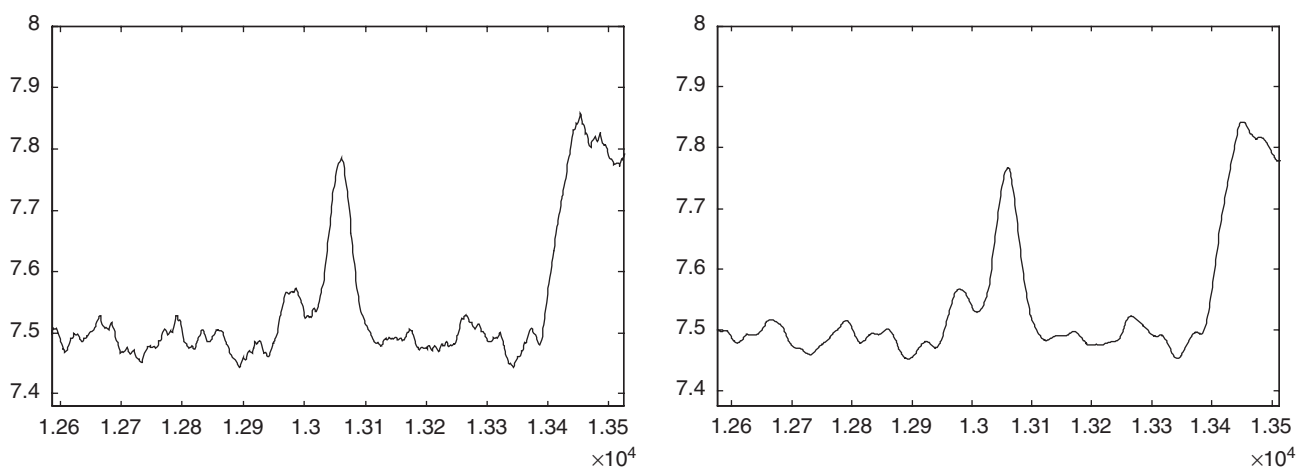


Figure 57.4 An example of smoothing. The original profile (left panel) demonstrates a high-frequency component. By averaging the signal locally, the high-frequency component is removed. Note that this results in a loss of information if this component carries useful content.

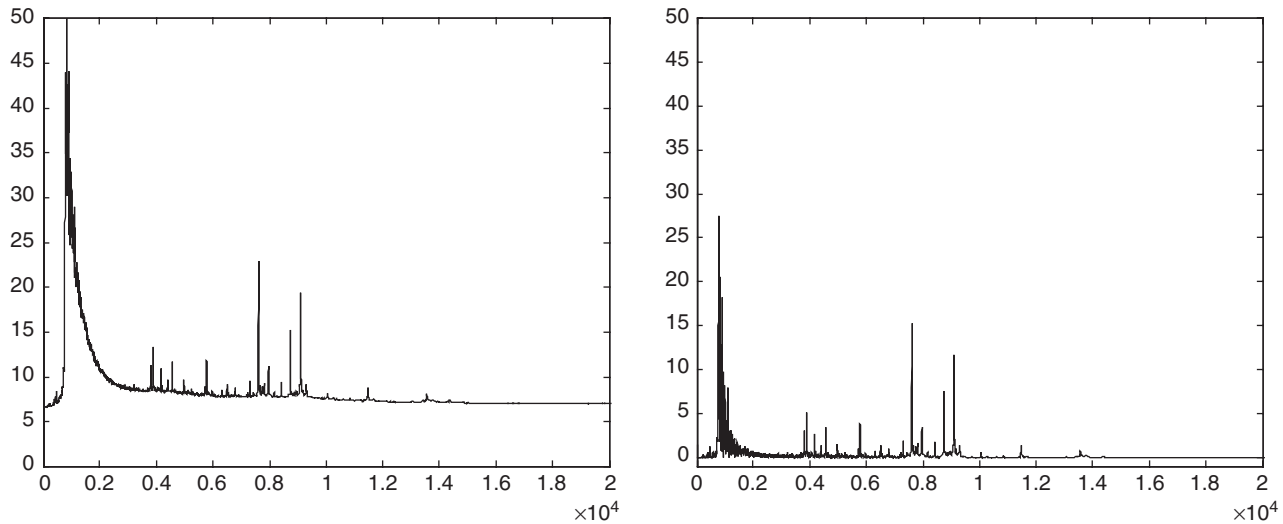


Figure 57.5 An example of baseline correction. Left panel: a profile with a baseline drift. Right panel: the corrected profile. The additive component in the signal is removed and the baseline is shifted to the zero intensity level.

BUILDING A PREDICTIVE MODEL

One of the objectives of SELDI-TOF MS data analysis is to build a predictive model that is able to determine the target condition (case or control, cancer or non-cancer) for a given patient's profile. The predictive model is built from a set of SELDI-TOF profiles (samples) assembled during the study. Each sample in the dataset is associated with a class label determining the target patient condition (case or control, cancer or non-cancer) we would like automatically to recognize. Our objective is to exploit the information in the data and to construct (or learn) a classifier model that is able to predict accurately the label of any new profile. Such a model can be then used to predict labels of new profile samples (diagnosis). The ultimate goal is to build (learn) the best possible classifier, i.e. a model that achieves the highest possible accuracy on future, yet to be seen, proteomic profile samples.

Many types of classification models exist. These include classic statistical models such as logistic regression (Kleinbaum, 1994), linear and quadratic discriminant analysis (Duda et al., 2000; Hastie et al., 2001), or modern statistical approaches such as *support vector machines* (Vapnik, 1995; Burges, 1998; Scholkopf and Smola, 2002). In general, the model defines a decision boundary – a surface in the high dimensional space of profile measurements – that attempts to separate case and control profiles in the best possible way. The left panel of Figure 57.6 illustrates a linear surface (a hyperplane) that lets us separate 193 of the 200 samples from the ovarian study using the intensity information in three profile positions (0.0735, 0.0786, 0.4153 m/z). However, note that in general, the perfect separation of two profiles via a linear surface using just three positions may not be possible. This is illustrated in the right panel of Figure 57.6 where

the linear surface allowing the perfect separation of case and control profiles does not exist. This scenario leads to sample misclassification – the assignment of incorrect profile labels to some samples relative to the decision boundary.

EVALUATION OF A MODEL

To evaluate the quality of a classification model, one must determine the ability of the model to generate accurate predictions of future unseen profile data. Obviously, since such data are unavailable, we can simulate this scenario by splitting the available data (obtained during the study) into a 'training' set and a 'testing' set. The training set is used to learn/construct the classifier[s]. The learning process adjusts the predictive model (classifier) so that the examples in the training set are classified with a high accuracy. The ability of the model to predict the case and control samples in the future is evaluated on the test set. Figure 57.7 illustrates the basic evaluation setup.

The quality of a binary (case versus control) classification model may be determined among many different metrics. For the purposes of this review, the classification models are evaluated using statistics computed from the confusion matrix, a two-by-two grid that represents the types and percentages of correct and incorrect classifications. A sample confusion matrix is shown in Figure 57.8. The following useful measures can be derived from the confusion matrix:

- Error (misclassification) rate: $E = FP + FN$
- Sensitivity (SN): $\frac{TP}{TP + FN}$

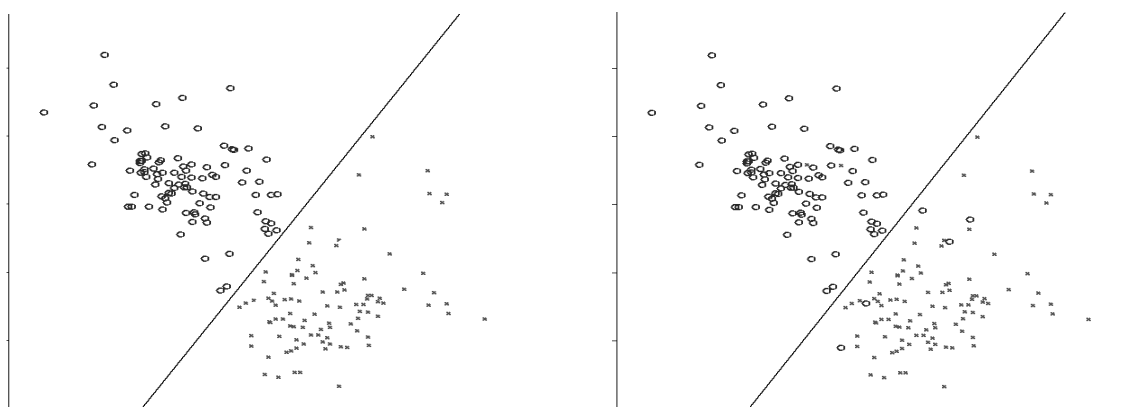


Figure 57.6 (Left panel) An example of a perfect separation of case (X) versus control (O) using a hyperplane in a 2-dimensional projection of the 3-dimensional space defined by m/z positions (0.0735, 0.0786, 0.4153) in the ovarian cancer study. Note that all controls samples are above the hyperplane while all cases are below. (Right panel) The perfect separation of the two groups does not exist. Some case and control samples appear on the opposite side of the hyperplane.

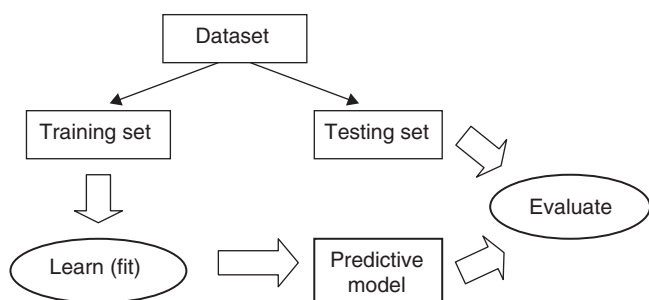


Figure 57.7 The basic evaluation setup. The dataset of samples (case and control profiles) is divided into the training and testing set. The training set is used to learn (fit) the predictive model and testing set serves as a surrogate sample of the future ‘unseen’ examples.

		Actual	
		Case	Control
Prediction	Case	TP 0.3	FP 0.1
	Control	FN 0.2	TN 0.4

Figure 57.8 Diagram of a confusion matrix. The four areas are labeled as follows:

- True positive (TP): the percentage of profiles which the classifier predicts correctly as belonging to the group of cases (cancerous)
- False positive (FP): the percentage of profiles which the classifier predicts as belonging to cases, but truly belong to controls (healthy)
- False negative (FN): the percentage of profiles which the classifier predicts as belonging to controls, but truly belong to cases
- True negatives (TN): the percentage of profiles which the classifier predicts correctly as belonging to the group of controls.

Note that the numbers should always sum to 1. An ideal classifier will have a value of 0 for both false negatives and false positives.

- Specificity (SP): $\frac{TN}{TN + FP}$
- Positive predictive value (PPV): $\frac{TP}{TP + FN}$
- Negative predictive value (NPV): $\frac{TN}{TN + FN}$

A confusion matrix and thus also the above measures can be generated for both the training and testing set. However, the evaluation of the testing set matters more; it is an indicator of how well the classification model generalizes to unseen data. Sensitivity and specificity measures on the testing set are very useful if we consider adopting the classification model for the purpose of clinical screening. Sensitivity reports the percentage of samples with a condition that were correctly classified as having the condition. Specificity, on the other hand, reports the percentage of samples without a condition that were correctly classified as not having the condition. Very high values of these statistics indicate a good screening test. Moreover, one type of error can be often reduced at the expense of the other error. This is very important if we care more about one type of error. For example, we may want to achieve a low number of patients incorrectly classified as suffering from the disease and care a bit less about missing a patient with the disease during screening.

In the simplest evaluation framework, statistics recorded in the confusion matrix are based on a single train/test set split. To eliminate a potential bias due to a lucky or an unlucky train/test split, the average of these statistics over multiple random splits is typically reported.

STRATEGIES FOR LEARNING A CLASSIFICATION MODEL

Classification models come in different guises. They may use a different decision boundary type or optimize slightly different learning criteria. Deeper analyses of many existing

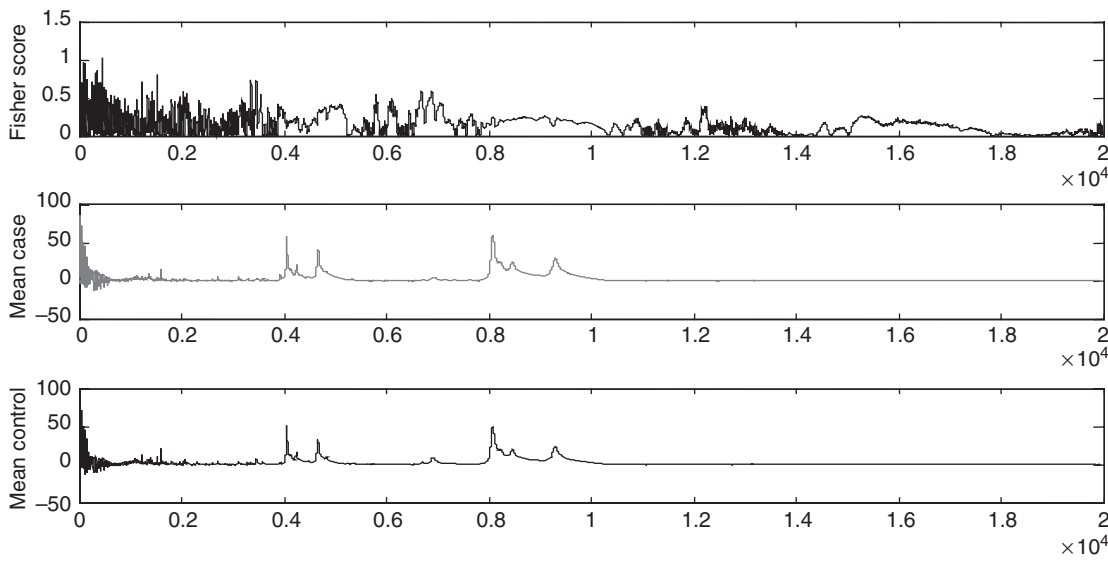


Figure 57.9 Differentially expressed features (profile positions). (Top panel) Fisher score values for each m/z position along the profile. (Middle panel) Mean of case profiles. (Lower panel) Mean of control profiles. Higher values of Fisher score indicate areas of the profile which are statistically likely to be good features.

models and their properties are beyond the scope of this work. To illustrate the main steps of the analysis we will focus on one such a model: the linear *support vector machine*. Briefly, the linear support vector machine models a linear decision boundary (a hyperplane) between different classes having the maximum ability to separate the groups of classes as illustrated in Figure 57.6.

Learning and predicting using a classification model is often hampered by the fact that the data are presented with a small number of samples, but have a naturally high dimensionality. In this case, there are only 200 profiles to study, yet the input vector of each profile contains 15 154 m/z values, or 'features'. When the number of features vastly outweighs the number of samples, the parameters of the classifier model are estimated with high variance. This makes it difficult to achieve a model that generalizes well to new examples. To alleviate this problem, we demonstrate the use of several types of *feature selection* approaches to convert the high-dimensional profile data into a low-dimensional dataset consisting of only a small number of m/z intensities. A hyperplane dependent on a smaller number of features is a more simplistic boundary and will generalize better to future unseen examples.

The primary goal of feature selection is to obtain a smaller set of features with high discriminative ability from a large set of inputs. To assess an ability of a profile position to discriminate between the case and the control a univariate statistical analysis can be performed. Several methods exist for evaluating the dispersion between cases and controls, including t-test, Fisher score or AUC score (Fisher, 1936; Hanley and McNeil, 1982; Hastie et al., 2001), etc. Figure 57.9 displays Fisher scores¹ for each position in the profile (top panel). Below, the mean case profile (middle

panel) and mean control profile (bottom panel) are shown. Higher values of the Fisher score indicate that the m/z position is likely represented differently in all case profiles versus all control profiles. These features can be more commonly referred to as *differentially expressed features*. Features with the highest Fisher score are very likely good feature candidates. These features may be memorized and used as *biomarkers* in screening additional profiles. Thus, selecting features is a short matter of computing the Fisher score of each feature (each m/z value) and then eliminating all but the top k features, where k is the desired quantity of 'left-over' features.

A univariate analysis of features ignores possible dependencies between the features. For example, the value of a feature may be highly correlated with values of other features. Because of this, many of the features may be redundant and their inclusion offers very little additional information towards building a predictive model. Figure 57.10 illustrates some of these problems. It shows 30 out of the top 100 positions that would be selected through Fisher score. We see that these 30 features accumulate within two areas related to two peak complexes. These groups of features are highly correlated and the amount of new discriminative information the duplicates add is relatively small. One way to alleviate this problem is by enforcing the maximum-allowed correlation (MAC) threshold on the features selected via univariate analysis (Hauskrecht et al., unpublished observations). This helps to ensure that the feature selection method increases its coverage of the signal and does not miss important information that might otherwise be ignored.

Focusing on individual positions in the profile may narrow the field of view. This is especially concerning if the discriminative signal is weak and noisy. One way to alleviate this problem is to look at aggregate features, or features that combine the information from many profile positions. The intuition is that the useful signal can be

¹The Fisher score is computed as $Fisher(i) = \frac{|\mu_i^+ - \mu_i^-|}{\sigma_i^+ + \sigma_i^-}$ where μ_i^\pm is the mean value for the i th feature in the positive or negative profiles, and σ_i^\pm is the standard deviation.

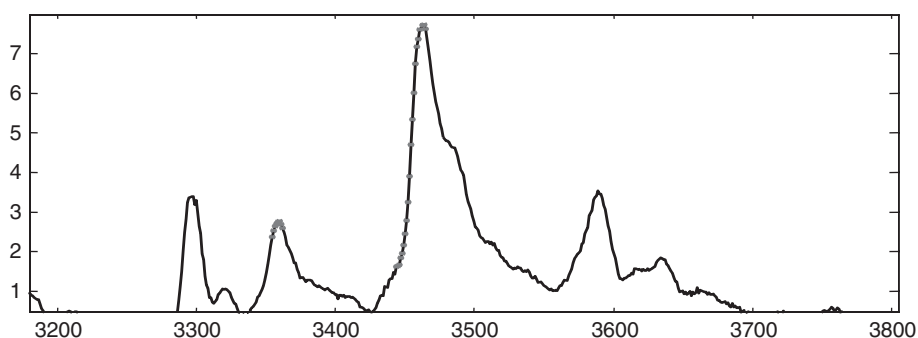


Figure 57.10 Thirty out of the top 100 positions selected by the Fisher score on the mean cancer profile. The positions (defined by the markers) accumulate on only two peak complexes and many features are highly correlated, thus carrying minimum additional discriminative information.

amplified over many related positions and thus it is less sensitive to random fluctuations/noise. An example of such a transformation is the principal component analysis (PCA) (Jolliffe, 1986). Intuitively, the PCA identifies orthogonal sets of correlated features and constructs aggregate features (or components) which are uncorrelated, yet are responsible for most of the variance in the original data. Retaining the variance is important; it is often helpful to explore parts of the data where the features spread out across a large space, which gives more room to find a decision boundary between classes. These methods often construct useful features due to their inherent ability to utilize the entire signal, rather than a limited number of positions.

Another popular solution that attempts to alleviate the problem of a noisy signal assumes that all relevant information is carried by peaks (Adam et al., 2002). The subsequent discriminative analysis is restricted to identified peaks only and/or their additional characteristics. Numerous versions of peak selection strategies exist (Adam et al., 2002). However, the utility and advantage of these strategies over other feature selection methods for the purpose of discriminant analysis has not been demonstrated.

RESULTS OF INTERPRETIVE ANALYSIS

The goal of this section is to illustrate the potential of the SELDI-TOF profiling technology on the analysis of the ovarian cancer data from April 2002 (Petricoin et al., 2002). The dataset² consists of 200 profiles samples: 100 samples represent cancer, 100 samples are controls. We rely on concepts and approaches discussed in previous sections and show that it is indeed possible to build predictive models that can classify with high accuracy test samples taken from the SELDI-TOF dataset.

The predictive model used in all our experiments is the support vector machine (see above). The model builds a linear decision boundary that separates cancer and control samples provided in the training set. In all experiments the model is learned using a limited number of features (5–25). We try different feature selection approaches. The process of finding a good set of features is a highly exploratory process and often remains the bottleneck of the discovery.

Table 57.1 Predictive statistics for the linear SVM model on the ovarian cancer dataset. The features are selected according to the Fisher score criterion. The maximum allowed correlation (MAC) threshold is 0.8. Test errors range in between 4 and 2.9 per cent. Sensitivities and specificities are between 94.9 and 97.6 per cent

No. of features	Testing error	Sensitivity	Specificity
5	0.0352	0.9764	0.953
10	0.0402	0.9698	0.9497
15	0.0406	0.9584	0.9604
20	0.0332	0.9641	0.9695
25	0.0299	0.9666	0.9736

Table 57.1 illustrates the quality of the predictive model learned using the top 5–25 *m/z* positions according to the Fisher score criteria. To remove highly correlated features we used the maximum allowed correlation (MAC) threshold of 0.8. The table shows the test errors (E), sensitivities (SN) and specificities (SP) for a different number of features. All statistics listed are averages over 40 different splits of the data into training and testing sets. This assures that the results are not biased due to a single lucky and unlucky train/test split. The split proportions are 70:30, i.e. 70 per cent of samples are assigned to the training set and 30 per cent to the test set.

Using the univariate statistical analysis approach as discussed above, the results are quite impressive. The best result occurs when the classifier is allowed to use the top 25 features selected by Fisher score. Under this condition, the classification model achieves 96.6 per cent sensitivity and 97.36 per cent specificity. On average, 2.99 per cent of the samples seen during the testing phase were misclassified. A different number of features used can show a tradeoff in the improvement of sensitivity or specificity. Note that sensitivity is highest when using only five features, yet specificity is highest when using 25.

Instead of narrowly examining a small number of individual positions, we can examine the effectiveness of aggregate features. Table 57.2 illustrates the performance statistics of our classification model using features constructed using PCA. The results are included over a range of 5 to 25 principal component features, which amplify patterns found in the profile signal. Again, the resulting statistics are averages over the same

²The ovarian cancer dataset is available at <http://ncifdaproteomics.com/>

Table 57.2 Predictive statistics for the linear SVM model on the ovarian cancer dataset. The features are constructed using principal component analysis (PCA). Test errors range between 19.9 and 8.9 per cent. Sensitivities and specificities range between 85.3 and 91.1 per cent

No. of features	Testing error	Sensitivity	Specificity
5	0.1992	0.8533	0.7477
10	0.1111	0.9161	0.8615
15	0.1078	0.8998	0.8846
20	0.0926	0.9038	0.911
25	0.0898	0.9087	0.9118

40 train/test splits as used in the univariate analysis, to allow for a more direct comparison of behavior.

The predictive performance of our classifier falls when used in conjunction with principal component features. The reason may be due to too many independencies between positions in the profile. Such a condition causes a problem for PCA, which attempts to find signal-wide relationships between multiple positions. If the signal-wide relationships that do exist are weak, then the benefits from using these features will be minimal. In addition, including many of these features to compensate for their weakness complicates the process of discovering biomarkers.

As a final example of feature selection, we illustrate the behavior of the aforementioned *peak selection* strategy. We refer to a peak as the local maximum over a region in the profile. The signal is smoothed before detecting the maxima on the mean case and control profiles. Peak positions from both mean profiles are then used as the primary set of features. Table 57.3 displays performance statistics of the above model using peak selection prior to selecting the top Fisher score features. Unfortunately, the performance was below what was achieved without the peak selection strategy (see Table 57.1).

Testing error is relatively high considering the results presented earlier in Table 57.1. The likely reason is that informative features are not only found on peaks, but in valleys as well. Another reason is that the particular behavior of the peak detection algorithm is not optimal. As mentioned before, there exist many methods for performing peak selection. Different criteria for selecting peaks will undoubtedly yield differing results.

The results presented above show that it is possible to learn predictive models that can achieve a very low classification error on SELDI-TOF samples. To support the significance of these results, in particular, the fact that the sample profiles carry useful discriminative signals, one may want to perform additional statistical validation. The goal of one such a test, the random class-permutation test, is to verify that the discriminative signal captured by the classifier model is unlikely to be the result of the random case versus control labeling. Figure 57.11 shows the result of the random permutation test for classifiers analyzed in Table 57.1. The figure plots the estimate of the mean test error one would obtain by learning the classifier

Table 57.3 Performance statistics of the linear SVM classifier after using peak detection. A MAC threshold of 0.8 was enforced before selecting the top 5–25 features using the Fisher score. Testing error ranges from 11.6 to 9.8 per cent, while sensitivity and specificity range from 85 to 93.4 per cent

No. of features	Testing error	Sensitivity	Specificity
5	0.1168	0.9152	0.8508
10	0.1049	0.9169	0.873
15	0.1033	0.9209	0.8722
20	0.0984	0.934	0.8689
25	0.1012	0.934	0.8631

on 5–25 features for randomly assigned class labels and estimates of 95 per cent and 99 per cent test error bounds. The estimates are obtained using 100 random class-label permutations of the original ovarian dataset. The results illustrate a large gap between classification errors achieved on the data and classification errors under the null (random class-label) hypothesis. This shows that our achieved error results are not a coincidence.

DISCUSSION

This review deals solely with clinical proteomics, but the analysis techniques reported are typically those that could be used in applications to other domains, including immunologic factors determined in the serum including cytokines, chemokines and antibodies or microarray data/transcription profiling of the peripheral blood. The profiles generated by SELDI-TOF MS are a rich source of information, which floats to the surface after careful analysis. Although there are many ways to analyze and evaluate proteomic profile data, a simple framework such as the one presented above serves as a foothold for future data analysis work.

Using proper feature selection techniques, proteomic profiling can be a valuable discovery tool for locating protein expression patterns in separate case and control populations. As seen above, by comparing expected generalization results on an unseen testing set, one can evaluate the performance of many feature selection strategies. The resulting classification models each contribute knowledge about the profiles, whether there is success or failure with subsequent test sets. In the case explored above, the peak selection strategy used was not effective due to important information being expressed in the 'valleys' of the profile. When these were taken into account, the predictive ability of the classification model is higher. When features were found to cluster among one another due to correlation in relation to proximity of m/z values, de-correlation became an important step in the process. Ultimately, the classification model was able to obtain a testing error of 3 per cent when using the top 25 m/z intensities ranked by Fisher score and having intra-correlation coefficients less than 0.8.

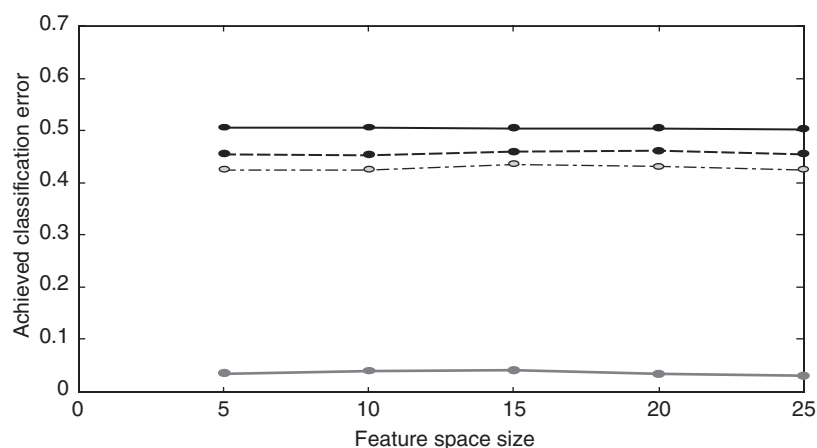


Figure 57.11 Permutation-based validation. The top solid line indicates the mean-achieved classification error (MACE) for the model under the null hypothesis: the class labels in the data are assigned to profiles randomly. The upper dashed line indicates the estimate of the upper 95th percentile of the test error statistic under the null hypothesis. Likewise the lower dashed line indicates the 99th percentile. The bottom solid line indicates the achieved classification error (ACE) using the original data labeling. Only the values for the points marked are computed.

The m/z values selected through the favored classifier play the most important role in clinical diagnostics. Databases of peptide masses are easily queried with a list of m/z values, which return possible protein sources for those peptide molecules. In this way it is possible to verify the statistically located biomarkers with the support of biological knowledge surrounding these indicators. Even if identifying the protein is not feasible, simply making a note of the m/z values is often enough. New patients can have their MS spectra generated and these m/z values checked in order to determine the efficacy a treatment will have, the progress a disease has made, or to determine composition of antibodies during an immune response.

Mass spectrometry proteomic profiling is certainly a viable technique for the discovery and detection of biomarkers. Advances in sample preparation and instrumentation are improving the usage of this technology. With continued application and development, the information derived from mass spectrometry studies will contribute substantially to what is currently known about the role, variation and relative abundance of multiple proteins in the setting of normality and disease.

The study of cell biology is, for a large part, dependent on solutions to many structural problems. Accurate structural characterization of protein peptides in femtomole levels is especially important to immunology and the study of proteomics. Surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF MS) is a technology which can assist biologists in studying the 'structures' of macromolecules in complex and varying situations; the ultimate goal is to utilize fully this technology for determining the function and abundance of peptides as they relate to health and disease. One important potential application of this technology includes identifying peptides/proteins that are associated with immune responses to diseased cells. Peptide or lipid antigens recognized by T cells in the context of MHC molecules serve as a means to identify relative changes or new proteins or pathogens residing within a cell, serving

as a basis for T cell recognition and effector function including the ability to lyse the cell presenting the nominal antigen. These antigens are presented as protein fragments which can be readily identified by the mass spectrometry technology (Castelli et al., 1995).

REFERENCES

- Adam, B.L., Qu, Y., Davis, J.W. et al. (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.
- Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining Knowl Discov* 2, 121–167.
- Castelli, C., Storkus, W.J. et al. (1995). Mass spectrometric identification of a naturally-processed melanoma peptide recognized by CD8+ cytotoxic T lymphocytes. *J Exp Med* 181, 363–368.
- Duda, R.O., Hart, P.E. and Stork, D.G. (2000). *Pattern classification*, 2nd edn. John Wiley and Sons. [**2]
- Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Ann Eugenics* 7, 79–188.
- Hanley, J. and McNeil, B. (1982). The meaning and use of the area under a receiver operating characteristic curve. *Diagnost Radiol* 143, 29–36.
- Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The elements of statistical learning*. New York: Springer-Verlag.
- Jolliffe, I.T. (1986). *Principal component analysis*. New York: Springer-Verlag.
- Kleinbaum, D.G. (1994). *Logistic regression: a self-learning text*. New York: Springer-Verlag.
- Patel, S. and Lyons-Weiler, J. (2004). caGEDA: a web application for the integrated analysis of global gene expression patterns in cancer. *Appl Bioinformat*, in press. [**3]
- Pawletz, C., Trock, B., Pennanen, M. et al. (2001). Proteomic patterns of nipple aspirate fluids obtained by SELDI-TOF. *Dis Markers* 17, 301–307.
- Petricoin, E.F., Ardekani, A.M. and Hitt, B.A. (2002). Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 359, 572–577.
- Petricoin, E.F. and Ornstein, D.K. (2002). Serum proteomic patterns for detection of prostate cancer. *J Natl Cancer Inst* 94, 1576–1578.

- Scholkopf, B. and Smola, A. (2002). Learning with kernels. MIT Press.[**4]
- Steel, L.F., Shumpert, D. and Trotter, M. (2003). A strategy for the comparative analysis of serum proteomes for the discovery of biomarkers for hepatocellular carcinoma. *Proteomics* 3, 601–609.
- Vapnik, V.N. (1995). The nature of statistical learning theory. New York: Springer-Verlag.
- Wadsworth, J.T., Somers, K. and Stack, B. (2004). Identification of patients with head and neck cancer using serum protein profiles. *Arch Otolaryngol Head Neck Surg* 130, 98–104.
- Watkins, B., Szaro, R., Shannon, B. et al. (2001). Detection of early stage cancer by serum protein analysis. *Am Lab* 32–36.[**5]
- Zeindl-Eberhart, E., Haraida, S. and Liebmann, S. (2004). Detection and identification of tumor-associated protein variants in human hepatocellular carcinomas. *Hepatology* 39, 540–549.
- Zhukov, T.A., Johanson, R.A., Cantor, A.B., Clark, R.A. and Tockman, M.S. (2003). Discovery of distinct protein profiles specific for lung tumors and pre-malignant lung lesions by SELDI mass spectrometry. *Lung Cancer* 40, 267–279.