ORIGINAL RESEARCH ARTICLE

Appl Bioinformatics 2005; 4 (4): 227-246
1175-5636/05/0004-0227/$34.95/0

Feature Selection for Classification of SELDI-TOF-MS Proteomic Profiles

Milos Hauskrecht,[1,2] *Richard Pelikan*,[1] *David E. Malehorn*,[2,3,4] *William L. Bigbee*,[2,3] *Michael T. Lotze*,[2,5] *Herbert J. Zeh III*,[2,5] *David C. Whitcomb*[6] and *James Lyons-Weiler*[2,7,8,9]

1 Department of Computer Science, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
2 University of Pittsburgh Cancer Institute, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
3 Clinical Proteomics Facility, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
4 Department of Pathology, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
5 Department of Surgery, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
6 Departments of Medicine, Cell Biology and Physiology, and Human Genetics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
7 Cancer Biomarkers Laboratory, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
8 Centers for Pathology and Oncology Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
9 Center for Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

Abstract

Background: Proteomic peptide profiling is an emerging technology harbouring great expectations to enable early detection, enhance diagnosis and more clearly define prognosis of many diseases. Although previous research work has illustrated the ability of proteomic data to discriminate between cases and controls, significantly less attention has been paid to the analysis of feature selection strategies that enable learning of such predictive models. Feature selection, in addition to classification, plays an important role in successful identification of proteomic biomarker panels.

Methods: We present a new, efficient, multivariate feature selection strategy that extracts useful feature panels directly from the high-throughput spectra. The strategy takes advantage of the characteristics of surface-enhanced laser desorption/ionisation time-of-flight mass spectrometry (SELDI-TOF-MS) profiles and enhances widely used univariate feature selection strategies with a heuristic based on multivariate de-correlation filtering. We analyse and compare two versions of the method: one in which all feature pairs must adhere to a maximum allowed correlation (MAC) threshold, and another in which the feature panel is built greedily by deciding among best univariate features at different MAC levels.

Results: The analysis and comparison of feature selection strategies was carried out experimentally on the pancreatic cancer dataset with 57 cancers and 59 controls from the University of Pittsburgh Cancer Institute, Pittsburgh, Pennsylvania, USA. The analysis was conducted in both the whole-profile and peak-only modes. The results clearly show the benefit of the new strategy over univariate feature selection methods in terms of improved classification performance.

Conclusion: Understanding the characteristics of the spectra allows us to better assess the relative importance of potential features in the diagnosis of cancer. Incorporation of these characteristics into feature selection strategies often leads to a more efficient data analysis as well as improved classification performance.

Proteomic profiling is an increasingly popular tool in the search for novel surrogate biomarkers for cancer diagnosis, prognosis and measures of response to therapy.[1-3] The greatest promise of proteomic profiling is the possibility of early detection of the disease using a relatively inexpensive and minimally invasive (serum) or completely noninvasive (urine) biospecimen. Hope also exists that different cancer types[4] and therapy-response

phenotypes might be reproducibly distinguished, thereby allowing earlier application and assessment of appropriate therapies.

The majority of the work on surface-enhanced laser desorption/ionisation time-of-flight mass spectrometry (SELDI-TOF-MS) analysis focuses on the classification problem (see, for example, Adam et al.,[3] Jones et al.[5] and Petricoin et al.[1]). The potential efficacy of the SELDI-TOF-MS serum protein profiling for cancer classification has been recently demonstrated in multiple studies, including human breast,[6,7] colon,[7] head and neck,[8] liver,[9] lung,[10,11] ovarian[1,12] and prostate cancer.[13-16] The studies describe diagnostic profile features and learning algorithms based on these features, which provide at least 80%, and in some cases >90%, classification accuracy between cancer cases and controls. From a clinical viewpoint, such positive results demonstrate the potential promise of this new bioassay technology.

The objective of the classification analysis is to build a predictive model to classify profile samples with the best possible generalisation performance. The success of the model-building process is measured using a relatively standard training and testing setup. In the training stage, the classifier is learned (or built) using a subset of samples obtained during the study. In the testing stage, the ability of the model to correctly predict the samples as diseased or healthy is tested on the independent subset of samples that were withheld from the training stage. The performance on the withheld data is used to approximate the accuracy of the model on future, yet to be seen, samples.

The learning task is especially challenging in the case of high-throughput proteomics data because of the dimensionality problem. Typically, the total number of cases available to train the classifier is small relative to the number of intensity measurements of each proteomic profile. The problem is of a statistical nature. The parameters of a classifier that uses readings at all profile positions cannot be reliably estimated using a small number of samples, and estimates are made with high variance. This prompts the development of feature reduction techniques that convert high-dimensional proteomic profiles to a small set of features that in turn can be input into a classifier.

There are many potential ways of defining and performing feature reduction. One of the classical solutions is the selection of a smaller subset of features derived from a large set of profile inputs. The classifier is then built using the reduced input set. Options for this approach range from finding positions in the profiles that exhibit statistically significant differences (as performed in many gene selection approaches to microarray analysis) to 'peak' selection algorithms,[17] which use only positions on profiles that correspond to peaks, and binning,[18] which uses every *n*th position.[19] An alternative approach to dimensionality reduction is to construct a small set of aggregate features such that every feature combines many inputs in the profile. Examples of such approaches include clustering,[20] principal component analysis (PCA)[21] and independent component analysis (ICA).[22,23]

The majority of published previous work on analysis of proteomic data has focused on one feature reduction/classifier approach. Although the results obtained in such a way clearly illustrate the ability of the proteomic profiles to differentiate between cases and controls, no answer has been provided to the question of which features are best suited for such a task. In contrast, we believe that some of the results and the means by which they are presented, especially in the earlier literature on proteomic profiling, provide an incomplete and somewhat misleading picture about the relevance of certain features.

The understanding of the importance of various features is crucial for obtaining robust (i.e. reproducible) classifiers and for finding reliable and accurate disease signatures. This is one of the main objectives of this study. Following this path, we present the results of the initial analysis of a number of simple feature reduction strategies on a pancreatic cancer dataset. Our intention is to build the understanding of the proteomic data and of the effects of various features on their ability to discriminate case versus control profiles. Since the meaning of features is defined in the context of the classification task, we evaluate the performance of feature reduction strategies indirectly by examining the quality of classifiers built upon generated features.

The pancreatic dataset used in this study was collected at the University of Pittsburgh Cancer Institute (UPCI), Pittsburgh, Pennsylvania, USA.[1] It consists of 116 profile samples, with 57 cases and 59 smoking-, age- and sex-matched controls. We start by examining the pancreatic data using relatively simple differential feature selection methods based on univariate analysis. Unfortunately, these methods promote features that lead to the best individual predictors; thus, the benefit of combining many of these features can be marginal. To alleviate this problem, we present a multivariate refinement of differential methods that is based on feature de-correlation. The idea of the approach is to ensure that none of the features selected is correlated with other features by more than some fixed correlation threshold. The method appears to be quite effective in the context of proteomic profiles with many highly correlated positions. We compare the refined method with a peak selection approach, where features are restricted to peak positions. Peak selection methods are quite popular among SELDI-TOF-MS researchers in particular, because peaks, as observed in profiles, are believed to represent discrete proteins or

1 The pancreatic data were provided courtesy of Herbert J. Zeh III, David C. Whitcomb and William L. Bigbee.

their fragments. We show that, in terms of performance, the differential feature selection methods with de-correlations compare favourably with performance of peak selection methods. The advantage of these methods is that they do not restrict their attention to peaks only; instead, the whole profile is searched for discriminative features. Although the results presented in this article do not give a definite answer to the problem of feature selection, we believe they provide new insights into feature selection approaches and thus help to narrow the search for clinically significant cancer signatures.

Data Source and Data Preprocessing

SELDI-TOF-MS

Ciphergen Biosystems Inc. (Fremont, CA, USA) SELDI-TOF-MS is used for the mass analysis of compounds such as proteins, peptides and nucleic acids (with masses ≤200 000Da) within solutions such as serum, urine, cell lysates or metabolites. Profiles may be determined for whole sera via 'neat spotting' or for fractionated samples, with profiles determined for each sample fraction. SELDI-TOF-MS operates by capturing compounds of interest on a chip (ProteinChip® Array, Ciphergen Biosystems Inc.). The surface of the chip possesses affinity characteristics such as ion exchange, hydrophobicity or antigen/antibody, which cause components of the mixture to selectively adsorb to the surface of the chip. Contaminants are removed by washing. After addition to the chip of 'energy absorbing molecules' (matrix), the remaining bound substances are analysed under high vacuum by laser desorption/ionisation time-of-flight mass spectrometry. The time of flight through the vacuum is converted to provide inferred molecular weight information using calibrated mass standards. An example of a profile is shown in figure 1.

Profile Inaccuracies

The profiles obtained through SELDI-TOF-MS are far from perfect. Multiple sources of variations affect the profiles. These include sample-to-sample variation in sample collection, variation in processing protocols,[24] variation in the instrument conditions over time, variation in the intensity of current and laser intensity fluctuations, and different magnitudes of the signal due to surface irregularities. All these affect our ability to analyse the profiles. In addition to systematic sources, if we analyse samples from multiple individuals then *natural biological variations* in sera are observed and detected. These show up as differences in intensity values or as the presence or absence of features in the profile. All sources of variations lead to serious challenges in interpretive analysis for the case versus control profile differentiation.

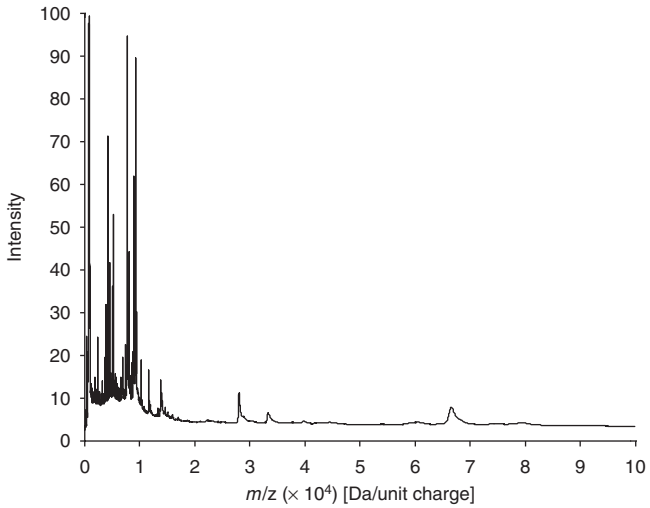

Fig. 1. A typical surface-enhanced laser desorption/ionisation time-of-flight (SELDI-TOF) profile plotting relative ion flux vs mass-to-charge (*m*/*z*) ratio. Note the relative abundance of species below 20 000Da.

Instrument variations manifest themselves as two types of errors: *intensity measurement errors* and *mass inaccuracy*. The intensity measurement error is due to an erroneous intensity reading of a certain mass-to-charge (*m*/*z*) ratio at the ion detector. Mass inaccuracy refers to the misalignment of readings for different *m*/*z* ratios. Figure 2 illustrates the scope of the problem by comparing two profiles obtained for the same pooled reference (quality assurance/quality control [QA/QC]) serum used by the Early Detection Research Network (EDRN) programme. The displayed profiles underwent external calibration by Ciphergen Biosystems Inc., and no other correction was applied.

Intensity Measurement Errors

Measurements are typically made with up to 400 laser 'shots' per spot, with analyses of the ionised molecules summed and averaged over these spectrograms. The intensity measurement error refers to the error of the intensity reading for a fixed *m*/*z* ratio. The signals in profiles exhibit, in addition to the random signal component, systematic additive and multiplicative error components.

The additive or *baseline error* refers to a systematic measurement error in which the baseline of all measurements on the profile differs from zero. This is apparent in figure 2a where the baseline measurements for the same serum differ from profile to profile. Moreover, we can also observe baseline drifts where the intensity of a baseline signal changes over regions within the profile. The multiplicative or *scaling error* refers to a profile-specific or region-specific error factor that affects the magnitude of the signal relative to the baseline. This is visible in figure 2b where the

intensity readings of the same sera vary greatly between the two profiles.

Mass Inaccuracy

All stochastic variations in profiles show up as differences in intensity readings. The mass accuracy (or variability in the physical location of *m*/*z* positions in the profile) is reported to be approximately 0.02Da if externally calibrated. This may appear small, but the mass drift (inaccuracy) can lead to serious challenges in the interpretive analysis of many samples. Any variation in *m*/*z* calibration translates, for all downstream analyses, into unwanted, potentially systematic, variation in intensity. This is evident in the comparison of QA/QC reference sera in figure 2c. Profiles suggest that even small amounts of a phase shift can lead to large differences in intensity readings and subsequent problems in interpretive analysis.

Preprocessing

Preprocessing includes steps taken to clean and modify data, with the expectation that most of the useful information content carried by the profiles is preserved. Typical preprocessing steps include *smoothing*, *rescaling*, *variance stabilisation*, *baseline correction* and *profile alignments*. Views on preprocessing steps differ widely. Many researchers prefer to work on unprocessed data, citing the risk of the loss of useful information or the introduction of systematic error during preprocessing (for example, see Baggerly et al.[25]). The tradeoffs are clear: significant preprocessing may eliminate our ability to capture biological variations and important clues in case versus control discrimination. On the other hand, unprocessed data can be so noisy that any useful information is degraded and difficult to work with.

Smoothing

Smoothing serves to eliminate a high-frequency component in the signal. Smoothing can be implemented through various techniques, the most popular of which involves fitting kernels (Gaussian, quadratic, etc.) to the signal. High-frequency variation is eliminated by local averaging of the signal. Note that smoothing is risky and may result in a loss of information if this high-frequency component carries real biological information.

Profile Rescaling

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 that seem to be on a different scale (see figure 2b). *Normalisation* or *rescaling* of these profiles allows us to compare them on the same scale. One simple strategy to standardise the profile can be conducted by dividing each position in a spectrum by the average intensity of the entire spectrum (equation 1):

$$x_i^* = x_i \frac{N}{TIC} \qquad \text{(Eq. 1)}$$

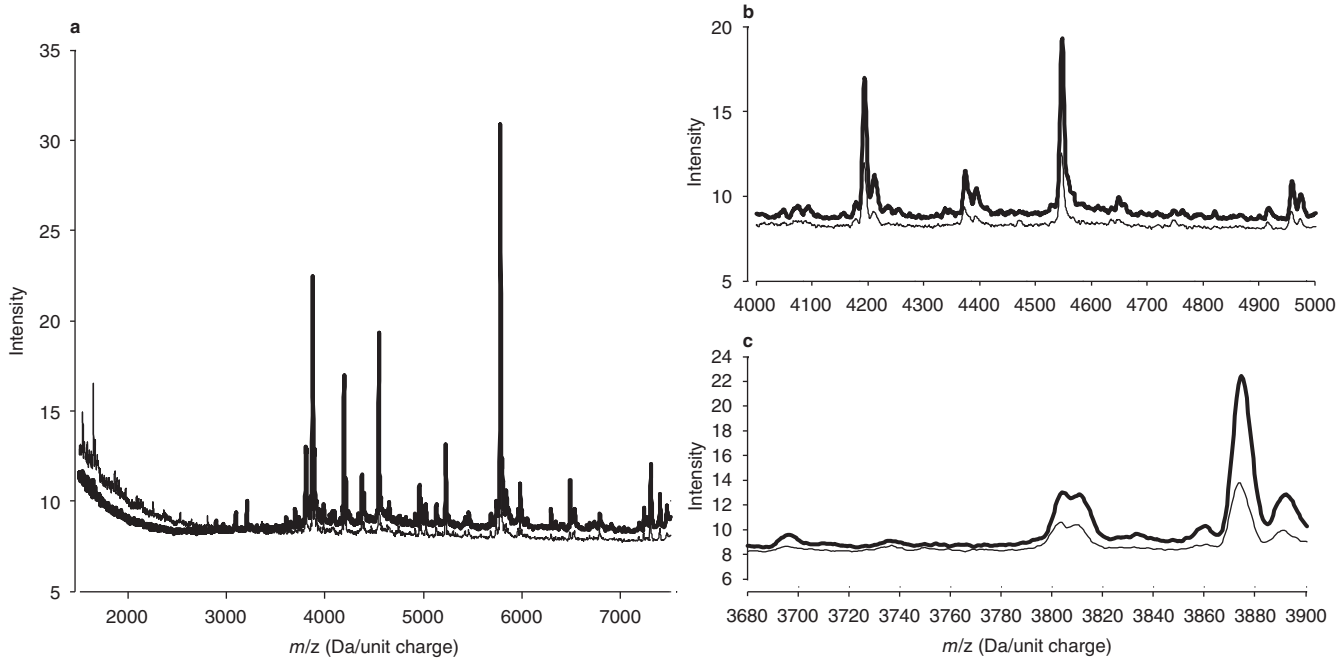

Fig. 2. Two surface-enhanced laser desorption/ionisation time-of-flight mass spectrometry (SELDI-TOF-MS) profiles obtained for the same pooled reference sera. The differences are apparent in the (**a**) baseline, (**b**) intensity measurements and (**c**) mass inaccuracies. ***m*/*z*** = mass-to-charge ratio.

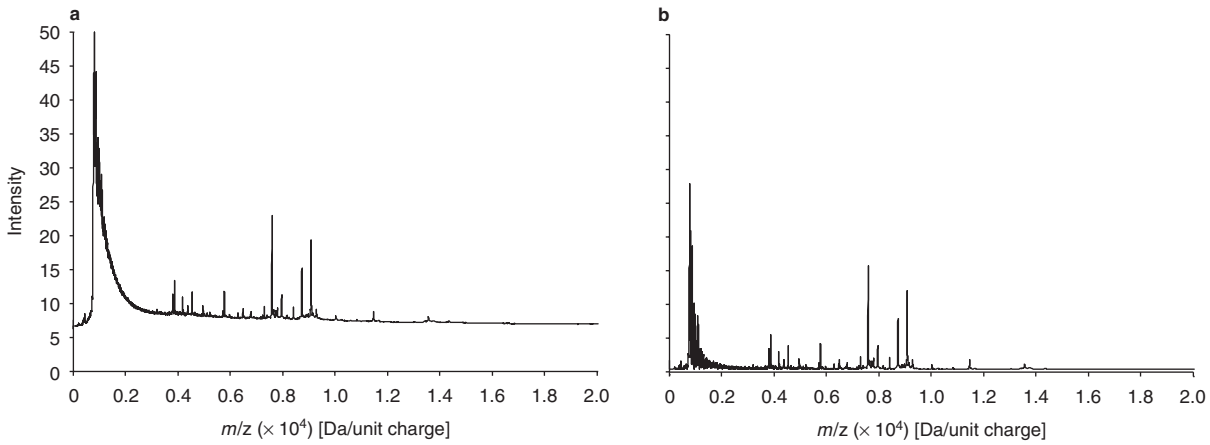

Fig. 3. Baseline correction on a surface-enhanced laser desorption/ionisation time-of-flight mass spectrometry (SELDI-TOF-MS) profile: (**a**) a profile with a baseline drift and (**b**) the corrected profile. The additive component in the signal is removed and the baseline is shifted to the zero intensity level. ***m/z*** = mass-to-charge ratio.

where x_i^* and x_i denote the new and previous intensity value, respectively, at position i; N denotes the number of intensity measurements; and *TIC* denotes the total ion current that is given by the sum of intensities under the profile curve. Another strategy is to correct the signal in the ith profile position[25] as (equation 2):

$$x_i^* = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad \text{(Eq. 2)}$$

where $x_{\min}$ and $x_{\max}$ denote the minimum and maximum intensity of the signal in the profile. This correction rescales the signal and transfers it into the [0,1] range. More complex normalisation approaches exist. For example, we recently developed a polynomial-based rescaling approach (Hauskrecht et al., unpublished data) that multiplies every profile using a low-order polynomial. The coefficients of the polynomial are found using linear regression methods. The detailed description and comparison of this method is outside of the scope of this article.

Variance Stabilising Transformations

One characteristic of SELDI-TOF-MS profiles is that the variance of the noise signal tends to be higher for higher intensity values. One possible technique to reduce the effect of such a noise component is to apply smoothing. Although this allows us to reduce the noise, the variability in the strength of the noise signal with regard to the intensity remains. A more methodical solution to correct for higher variance of the noise component in response to higher intensity values is to apply *variance stabilising transformations*,[26] such as square-root, cube-root[2] or logarithmic transforms.[19] By applying the transformation to intensity readings within the profile, the multiplicative noise component is stabilised.

Baseline Correction

The goal of *baseline correction* is to reduce the additive bias in the intensity signal (figure 2a). Figure 3 illustrates the process of baseline correction on a profile. The method removed the additive component in the signal and brought the baseline to zero. The baseline algorithm used in figure 3 uses local minima-based correction, where the baseline is defined by the minimum value over the local window of a fixed width. To account for what appears to be an exponentially decreasing baseline, the baseline correction is restricted to a non-increasing function after the local minimum has reached its highest value.

One issue that arises in the context of baseline correction is that the noise on the intensity measurement appears to be strongly correlated with the magnitude of the measurements; consequently, this means that any baseline correction results in the loss of information. Thus, any signal rescaling or transformation (e.g. cube-root transform) should be performed before the baseline is corrected.

Profile Alignment

The mass inaccuracy problem (see figure 2c) can be resolved through *profile alignment* methods. A number of strategies exist for performing profile alignment. One option is to define a reference profile in terms of a set of established biomarkers that are easily identifiable in every profile (internal calibration). Another approach is to include indicator peptides in the serum to purposely populate the profile with peaks to be expected at certain *m/z* values

2 The cube-root transform for SELDI-TOF-MS profiles was suggested to us by Jeffrey Morris (personal communication) from the University of Texas MD Anderson Cancer Center, Houston, Texas, USA.

(external calibration). The intensity readings between these peaks could then be stretched or shrunk along the *m/z* axis appropriately. Unfortunately, because of the locality of *m/z* errors, this approach would require the addition of several thousand peptides to the serum in order to properly recapture the information lost through mass inaccuracy.

More general profile alignment methods, particularly various incarnations of the time-warping algorithm,[27] attempt to realign two profiles by trying to minimise the differences in their signals. The algorithms rely on the dynamic programming paradigm. Unfortunately the memory and computational requirements of the method are quadratic in the length of the profile, which in the case of SELDI-TOF-MS consists of ~60 000 positions. Constraints on the maximum allowed warp shift[28] can alleviate the memory and computational problem to some degree. Alternatives to time warping via dynamic programming (with possible profile deletions and insertions) are parametric or semi-parametric time-warping methods.[29,30] In such a case, the warping function is restricted to a low-order polynomial, and the parameters of the polynomial are fit iteratively via regression by minimising the sum of the squared distance metric.

Preprocessing of the Pancreatic Cancer Dataset from the University of Pittsburgh Cancer Institute

The pancreatic cancer dataset analysed in this study was collected at the UPCI and includes 57 preoperative cancer cases and 59 age-, sex- and smoking history-matched controls. The serum samples were denatured and processed in duplicate on a single type (IMAC3-Cu) ProteinChip® Array. The IMAC3-Cu surface type chosen, which selectively retains metal-binding peptides/proteins, has been previously shown[31] to provide reproducible and feature-rich protein profiles from human serum. Serum samples were processed using robust procedures employed for an ongoing multisite validation of a SELDI-TOF-MS-based test for prostate cancer,[32,33] of which UPCI is a member, and all steps were conducted on a Biomek® 2000 liquid-handling robotic workstation (Beckman Instruments, Inc., Fullerton, CA, USA). Whole serum samples were denatured by mixing 20μL of serum with 30μL of 8 mol/L urea/1% 3-[(cholamidopropyl)dimethylammonio]-1-propanesulfonic acid (CHAPS) in phosphate-buffered saline (PBS) in a 96-well microtitre plate and incubating 30 min with shaking at 4°C. Samples were diluted by addition of 100μL of 1 mol/L urea/0.12% CHAPS in PBS, then performing a serial 1 **:** 5 dilution in PBS. From the final dilution, 100μL aliquots were reacted with a single spot on IMAC3-Cu ProteinChip® Arrays, which were preloaded with copper sulphate. Samples were applied in a blinded layout of case and control samples, along with one spot per ProteinChip® of a pooled reference serum sample for quality assurance. After 30 min incubation with shaking at room temperature, the chips were washed twice with PBS, rinsed twice with high-performance liquid chromatography (HPLC)-purity water and air-dried. Prior to mass spectrometry, two 1μL aliquots of a subsaturated solution of sinapinic acid were added to each spot, with applications separated by a 5-min drying time. The ProteinChip® arrays were read in a PBSIIc mass spectrometer (Ciphergen Biosystems Inc.) using positive ion mode, with time delay focusing, from 0 to 100 kDa. Mass calibration was performed externally (Ciphergen Biosystems Inc.) using a mixture of seven peptide species from 1 to 7 kDa.

The data preprocessing protocol was fixed and included variance stabilisation, baseline correction, smoothing and alignment. We performed a cube-root signal transformation to stabilise variance in the data. Following this, baseline correction was performed using an in-house baseline procedure that corrects the signal according to the local signal minimum. A window size of 200 *m/z* positions was used as a default to detect the minimum. We used Gaussian kernel smoothing to lightly smooth the signal. Finally, we performed a peak-based alignment by choosing peaks within the mean profile and warping them to fit the local characteristics of this reference profile.

Classification of Proteomic Profiles

The primary objective of proteomic profile data analysis is to build a predictive model that is able to determine the target condition (case or control) for a given patient's profile. The predictive classification model is built from a set of SELDI-TOF-MS 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) we would like to automatically recognise.

More formally, let D be a set of data pairs $\{<x_1, y_1>, <x_2, y_2>, \ldots, <x_n, y_n>\}$, where x_i denotes inputs and y_i their designated outputs. In the case of proteomic profiles, x_i corresponds to profile readings (a vector of *m/z* intensity values) and y_i to the class label: case or control (cancer or non-cancer). The objective is to build a predictive model f: $X \rightarrow Y$ that maps inputs (profiles) to outputs (labels) such that the mapping achieves high accuracy on future, yet to be seen, profiles. The classification (prediction) refers to the process of applying the learned model f: $X \rightarrow Y$ to profiles and assigning the output label for them.

Classifier Models

Many classifier models and learning approaches have been developed and are available for these classification tasks. Their

common property is that they represent the mapping between inputs and outputs. For example, some classifiers including CART® [34] and C4.5[35] extract classification rules in terms of decision trees. Some methods, including logistic regression,[36] determine the output by a learning set of parameters used to weight individual inputs. Other examples include support vector machines (SVMs),[37-39] the naive Bayes classifier[40,41] and multilayer neural networks.[42-44]

In general, classification models attempt to partition a high-dimensional space of profile measurements ($\boldsymbol{x}$), such that the case and control profiles fall into distinct regions. Many existing models, such as logistic regression or the SVM, achieve the partitioning by defining a linear decision boundary: a hyperplane that separates a high-dimensional input space $\boldsymbol{x}$ into two subspaces. Different models may use different optimisation criteria. For example, the SVM[37-39] is a technique that computes a decision boundary between two classes by restricting its attention only to the samples (support vectors) that are most critical for separating the two groups. In our case, the decision boundary is a hyperplane that is maximally distant from the support vectors on either side of the hyperplane. The hyperplane is defined by the equality (equation 3):

$$\boldsymbol{w}^T \boldsymbol{x} + w_0 = 0 \quad \text{(Eq. 3)}$$

with parameters $\boldsymbol{w}$ and w_0, where w_0 is the distance between the support vectors of each class, and $\boldsymbol{w}$ is the normal to the hyperplane.

The parameters of the model may be learned through quadratic optimisation with Lagrange parameters.[39] Then, the decision boundary is given by (equation 4):

$$\hat{\boldsymbol{w}}^T \boldsymbol{x} + w_0 = \sum_{i \in SV} \hat{\alpha}_i y_i (\boldsymbol{x}_i^T \boldsymbol{x}) + w_0 \quad \text{(Eq. 4)}$$

where $\hat{\alpha}_i$ are Lagrange parameters obtained through the optimisation process and y_i represents the class label for x_i with two possible values, −1 or 1. Note that only samples that correspond to support vectors (SV) define the hyperplane boundary, to which $\hat{\boldsymbol{w}}$ is normal. The decision $\hat{y}$ made by the classifier for a new input $\boldsymbol{x}$ is given by (equation 5):

$$\hat{y} = sign\left[\sum_{i \in SV} \alpha_i y_i (\boldsymbol{x}_i^T \boldsymbol{x}) + w_0\right] \quad \text{(Eq. 5)}$$

which corresponds to the side of the hyperplane on which the datapoint occurs, either positive or negative. The choice of the separating hyperplane in the SVM algorithm incorporates regularisation effects,[37] which makes it less susceptible to overfitting.

Evaluation of Classifier Methods

Our objective is to obtain models that achieve accurate predictions on future profiles. Since these examples are unobtainable, the ability of a classifier model f to generalise to such data is analysed by splitting the available data into two subsets: a training set and a test set. The training set consists of profile samples used to pick the features and learn the model. The test set consists of profile samples withheld from the learning stage that are used to approximate the ability of the classifier to correctly predict future, yet to be seen, data.

The complete performance picture is given by the confusion matrix that represents the percentages of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) results. Secondary measures can be derived from the confusion matrix and include the following:

Error rate (E): $FP + FN$
Sensitivity (SN): $TP/(TP + FN)$
Specificity (SP): $TN/(TN + FP)$
Positive predictive value (PPV): $TP/(TP + FP)$
Negative predictive value (NPV): $TN/(TN + FN)$.

These performance measures can be computed for both the training set and test set. Test set results are more important since they testify about how the classifier generalises to future data. However, the differences in training and testing performance statistics are still important and carry useful information. For example, a large separation between training and test errors is a sign of high variance of the estimates of the model parameters and indicates potential overfitting of the model.

Pitfalls and Solutions

Existence of a Perfect Classifier

Because of intrinsic stochasticity in the data and its source, it may be impossible to obtain a perfect classifier with zero expected error. For example, a noisy data source may influence the features used to discriminate a case and a control so that the two samples look alike and it is impossible to distinguish between them. In general, we should not expect perfect test errors in stochastic environments.

Lucky Train/Test Splits

Seeing a small test error on a single training/test split is a good sign, yet it is not always a reason for optimism. In general, it is not very hard to obtain a test result that appears to be very good if we look at many different data splits and pick the best result. For example, different training/test data splits on completely random data (with expected error of 50%) can produce different test errors, and a few among them may even obtain, by chance, an error value

equal or close to zero. So even if the data given to us are completely random and we cannot learn anything, it is possible to occasionally get low test errors by being lucky on a single training/test split.

Approximation of the Generalisation Performance

To eliminate a possible bias due to a lucky or an unlucky training/test split we can asses the quality of a classifier by learning and testing its performance on multiple (typically random) training/test splits and by averaging its predictive performance (in terms of test errors) on these splits. Cross-validation setups, such as random subsampling, *n*-fold, or leave-one-out validation can be applied and used to average the test error over multiple data splits.

Choosing the Best Classification Model

If we want to search among multiple classification models and find the best model, the choice of the best should be made exclusively during the training stage. The selection of the best classification model based on the test error is an unreliable and biased estimate of the selected model's generalisation performance. The reason is that such a test error does not report on the performance of the best model itself; it reports on the minimum test error statistic defined by all classifier models under consideration. In other words, when predicting future (unseen) examples, we do not know the best model to apply ahead of time. Thus, the correct model selection should be performed on the training set, and internal cross-validation should be used to pick the best alternative.

Permutation-Based Validation of Classification

The evaluation results can be strengthened through additional statistical validation tests.[45] In particular, one may ask whether the discriminative signal picked by our model in proteomic profiles is not the result of randomness. The significance of the result can be statistically validated by the random permutation test.[46,47] The random permutation test gives a non-parametric method to estimate the probability distribution of the statistics under the null hypothesis. In our case, the null hypothesis assumes that the relationship between the data and the labels (case or control) cannot be learned reliably by our feature-reduction/classifier model. Our goal is to reject this null hypothesis. The advantage of the test is that it can be applied to validate the significance of any classification model and the discriminatory signal detected therein.[48]

Receiver Operating Characteristic (ROC) Curves and Area Under the ROC Curve

The evaluation measures, discussed in the section titled Evaluation of Classifier Methods, are appropriate indicators of a learning model's performance under a 0-1 loss condition that reflects the situation in which type 1 and type 2 errors (FP and FN) carry approximately the same weight. However, many problems require a different loss function in which misclassifications of case and control use different weights. The ability of a binary classifier to incorporate a different loss function for the problem at hand can be captured and examined independent of the loss function in terms of the receiver operating characteristic (ROC). The separation of the two classes with different proportions of misclassification errors is measured and summarised using the area under the ROC curve (AUC) score.[49]

Multivariate Feature Selection Strategies

The learning and prediction steps are especially challenging in the case of high-throughput proteomics data because of the dimensionality problem: the total number of cases used to train the classifier is small compared with the dimensionality of each input vector. For example, the pancreatic cancer dataset used in this study consists of 116 profile samples, each with 60 264 *m*/*z* value measurements. Parameters associated with classifiers that rely on very large numbers of inputs cannot be reliably estimated because some are likely to attain discrimination power by chance. Moreover, when the number of individual measurements vastly outweighs the number of samples, parameter estimates are made with high variance. This prompts the development of feature reduction techniques that convert high-dimensional input data into a small set of highly discriminative features. Using such features leads to a less complex classifier whose parameters can be then estimated more reliably.

Ideally we want to identify a small set of features that let us separate with high accuracy the profiles in the two groups. If it is successful, the features would define the biomarkers. Moreover, if features correspond to *m*/*z* positions in the profile, we are provided with more information on the *m*/*z* class of peptides or protein complexes responsible for the differences. This can be of tremendous importance for the biochemical identification of the peptide/protein features, with implication for understanding the mechanism of the disease and the development of new therapeutic targets. In general, the process of identifying a small set of good and reliable features out of many possible options is a very challenging task and requires exploratory work.

The primary goal of this work is to explore and evaluate a number of multivariate feature selection approaches on the SELDI-TOF-MS proteomics data. In the simplest case, features correspond to *m*/*z* profile positions and their intensity readings. More complex features can characterise aggregate properties of multiple *m*/*z* positions or complete profiles. Examples of such

features include the slope or the energy of a peak in a certain *m/z* range or principal components of the profile. In the following subsections, we examine a set of feature selection strategies used to learn predictive models for the pancreatic cancer dataset collected at the UPCI.

The quality of features and feature reduction strategies cannot be reliably evaluated independent of the classification problem. Good features arise in the context of a classifier model and the accuracy of its predictions. Because of this close coupling, we propose to evaluate features indirectly in combination with one classifier model: the linear SVM classifier (see subsection titled Classifier Models).[3] Although more complex classifiers such as kernel-based SVMs[39] or decision trees[16] can be and have been applied, we believe the analysis based on simple classifiers can provide us with substantial information and enable us to derive insights about the characteristics of profiles and features that are important for discriminatory tasks. Fixing a classifier model applied to each of the feature strategies allows us to observe and evaluate the intrinsic value of each strategy in the context of such classifiers.

Feature Selection through Univariate Differential Expression

Choosing a good set of features is a challenging process. Searching of the complete space of feature subsets is an intractable task: there are *n-choose-k* different feature subsets of size *k* among *n* features. To alleviate this problem, the majority of practical feature selection methods choose features according to various simpler statistical criteria or heuristics. The most common approach (used frequently in microarray data analysis) is to evaluate the potential of every feature to discriminate profile samples individually through univariate statistical analysis and related univariate measures.

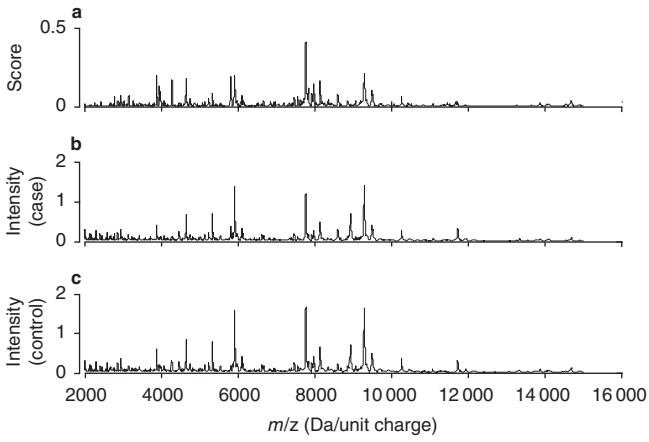

Fig. 4. Statgram for the SAM (significance analysis of microarrays) score and the pancreatic cancer dataset. (**a**) SAM score values for each mass-to-charge ratio (*m/z*) position along the profile. (**b**) Mean of case profiles. (**c**) Mean of control profiles. Positions with a high value of the SAM score are likely to be represented differently in case and control profiles. As visible in the figure, the positions with the highest SAM score exhibit a noticeable difference between means of the two sample groups.

A large number of univariate methods exist for determining the potential of a feature to differentiate between case and control. Examples include criteria or feature rankings based on the Fisher score,[51] scores based on the t-test,[4,52] Mann-Whitney U test,[10] AUC,[15,53] mutual information score[54,55] and many others. Any of the above scores can be used to define a relative order of profile positions in terms of their discriminative power or to filter out positions that violate some minimum score threshold.

An objective of feature reduction is to identify a small number of features with high discriminative power. Defining the feature selection criteria for high-dimensional feature spaces in terms of a single threshold criterion (such as the p-value threshold) may suffer from a large number of false positive identifications due to multiple comparisons. Simple multiple comparison corrections, such as the Bonferroni correction, attempt to quantify and control the risk of erroneously selecting at least one feature. Unfortunately, these methods are too conservative, especially for high-dimensional data sources. These problems can be alleviated by exploiting methods that provide estimates and bounds of the *false discovery rate*.[56] The false discovery rate method controls the fraction of false positives over the total number of positives.

In this work, instead of studying the effect of multiple comparisons on the number and quality of identified features we focus on relative ordering of univariate features. Figure 4 shows a statgram for the univariate SAM (significance analysis of microarrays) score.[57] The score is based on the statistic proposed for the analysis of microarray data, and it lets us order the features according to their ability to discriminate the case and control samples. The statistic is defined as the expression difference between the two groups normalised with respect to deviation of data (equation 6):

$$d(i) = \left| \frac{\bar{x}_+(i) - \bar{x}_-(i)}{s(i) + s_0} \right| \qquad \text{(Eq. 6)}$$

where $\bar{x}_+(i)$ and $\bar{x}_-(i)$ are the means of the case and control groups, and $s(i)$ is the 'feature-specific scatter' (equation 7):

3 To learn the SVM model, we use an iterative optimisation algorithm described by Mangasarian and Musicant.[50]

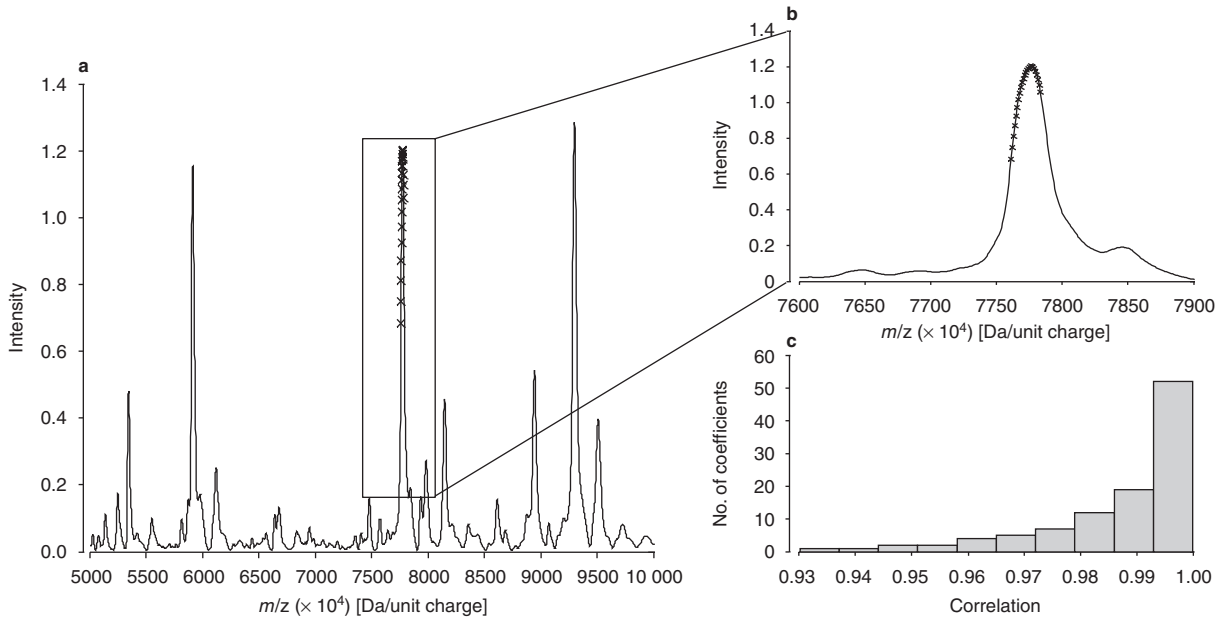

Fig. 5. (**a**) The positions of the top 15 SAM (significance analysis of microarrays) score positions on the mean control profile. All features occur on the same peak complex. (**b**) This area magnified. (**c**) A histogram plot of all correlation coefficients of the top 15 features. All correlation coefficients are high (>0.93). ***m/z*** = mass-to-charge ratio.

$$s(i) = \sqrt{\frac{(1/n_1)+(1/n_2)}{(n_1+n_2-2)}\left[\sum_{j=1}^{n_1}\left(x_j(i)-\bar{x}_+(i)\right)^2+\sum_{j=1}^{n_2}\left(x_j(i)-\bar{x}_-(i)\right)^2\right]} \quad \text{(Eq. 7)}$$

where n_1 and n_2 define the number of samples in the case and control groups. The coefficient s_0 lets us control the variance in $d(i)$. We use the default value of $s_0 = 1$ for all of our experiments.

By analysing figure 4, we see that many of the peak regions with high scores manifest visible differences in intensity expressions for case and control profiles. However, these differences are not deterministic; not all case and control samples can be distinguished using them. Thus, aggregate differences as captured by mean case and control profiles in figure 4 are often clearer.

The similar univariate scores tend to aggregate locally. Thus, the highest scores often include multiple profile positions in the same region. This is illustrated in figure 5a, which shows the top 15 SAM score positions in the profile. The result is not surprising, because it captures the nature of a typical proteomic profile: positions (especially those on peaks) are highly auto-correlated locally. Existing autocorrelations can be explained by a combination of a number of averaging and noise effects: (a) measurements of multiple laser shots are averaged; (b) a peak signature is not an ideal Dirac impulse, instead it is dispersed; and (c) positions of peaks are misaligned because of imperfect calibration, and also because profiles are smoothed using local averaging in the preprocessing stage.

De-Correlation Enhancement of Univariate Scoring Methods

Univariate statistical methods assess the ability of an individual feature to discriminate between the cases and controls. However, we are interested in obtaining a set of features with the best combined discriminative power. To address this problem, we choose and apply a simple refinement of the feature selection methods that builds upon the results of the correlation analysis. In particular, in addition to picking the best possible discriminative features as measured by the univariate score, we restrict the choice to only features that are intercorrelated by less than a maximum allowed correlation (MAC) threshold. Figure 6 shows the positions of the top 15 SAM scores with MAC = 0.6 on the mean cancer and mean control profiles.

One of the consequences of correlation-based filtering is that features are less likely to aggregate in the same region and along the same peaks. Note the differences between the positions of the top 15 features in figure 5 and figure 6. Without the de-correlation, all top 15 positions fall onto the same peak complex (figure 5).

The benefit of feature de-correlation is 2-fold. First, by eliminating close feature replicas we are able to select other important discriminatory features, thus improving our ability to discriminate well between cases and controls. Second, the replicate elimina-

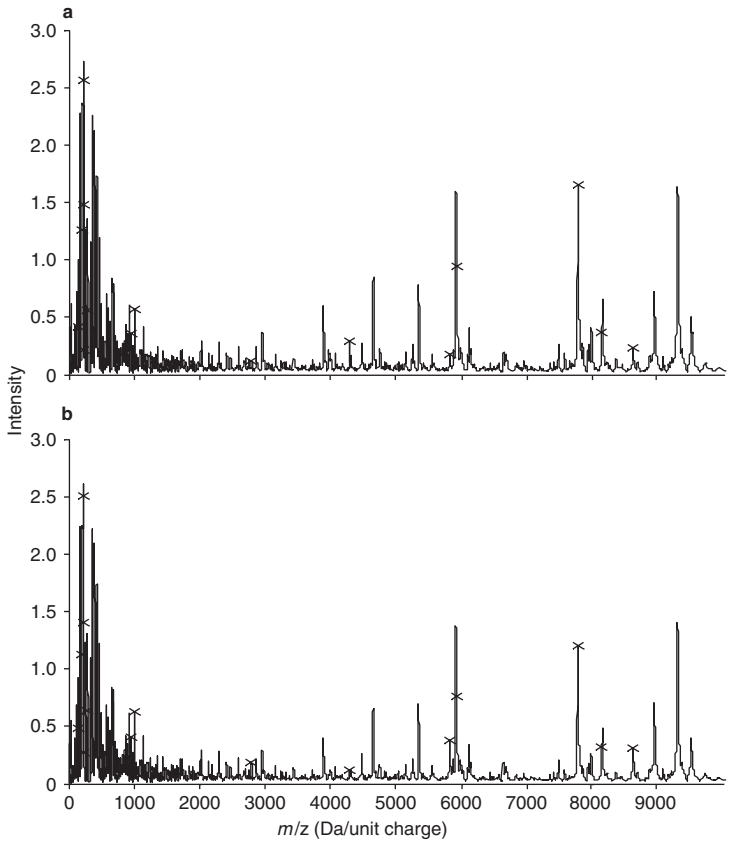

Fig. 6. Positions of the 15 highest-ranking SAM (significance analysis of microarrays) features (crosses on graphs) with the maximum allowed correlation of 0.6: (**a**) mean control profile and (**b**) mean cancer profile. The selected features (profile positions) are more distributed across the entire profile, due to the effect of de-correlation. ***m*/*z*** = mass-to-charge ratio.

tions prevent classification model overfitting, especially in models that are highly sensitive to correlated inputs.

To illustrate the effect of de-correlation on the classification accuracy, we consider a linear SVM classifier with features selected based on the SAM score and restricted by different MAC options. Figure 7a and table I (top half, labelled SAM) show the results of the experiments for the top 1–25 SAM score features under MAC thresholds of 0.6, 0.8 and 1.0 (no de-correlation). The table lists four evaluation statistics: test error, 95% confidence bounds on test error, sensitivity and specificity. The associated

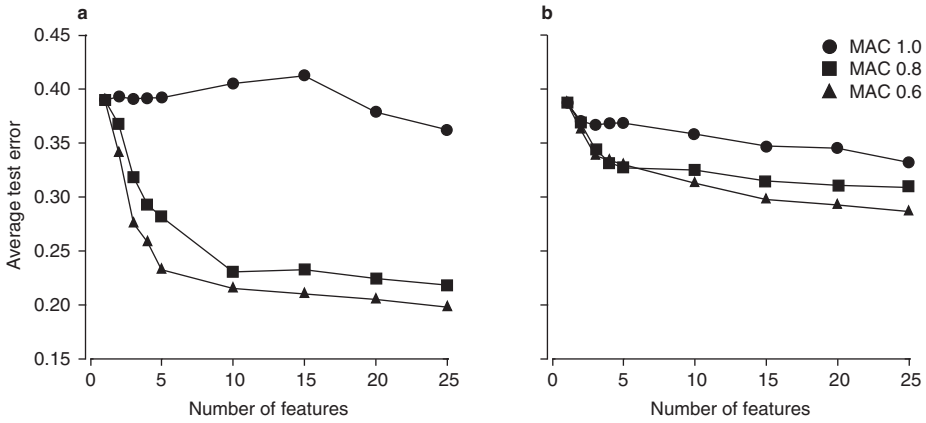

Fig. 7. The effect of maximum allowed correlation (MAC) thresholds on test errors. (**a**) Feature selection with the SAM (significance analysis of microarrays) score. (**b**) Feature selection with the Wilcoxon rank-sum score. Full performance statistics are given in table I. Test errors are given for varied number of features and for different de-correlation (MAC) thresholds.

Table I. Classification statistics for univariate feature scoring methods (SAM [significance analysis of microarrays], and Wilcoxon rank-sum test) with de-correlations and the linear support vector machine classifier. Three maximum allowed correlation (MAC) thresholds (0.6, 0.8 and 1.0) are compared for a varied number of features (1–25)[a]

Method used and no. of features	MAC = 1.0				MAC = 0.8				MAC = 0.6			
	test	95% CI	sens	spec	test	95% CI	sens	spec	test	95% CI	sens	spec
SAM												
1	0.3903	±0.0199	0.0641	0.5736	0.3903	±0.0199	0.0641	0.5736	0.3903	±0.0199	0.5736	0.6466
2	0.3931	±0.0210	0.0679	0.5722	0.3674	±0.0265	0.0857	0.586	0.341	±0.0312	0.6217	0.6971
3	0.391	±0.0198	0.064	0.575	0.3187	±0.0335	0.1082	0.6341	0.2757	±0.0311	0.6864	0.763
4	0.3917	±0.0212	0.0686	0.575	0.2931	±0.0316	0.102	0.6726	0.2576	±0.0264	0.7153	0.77
5	0.3924	±0.0210	0.0678	0.575	0.2812	±0.0337	0.1088	0.6891	0.2319	±0.0222	0.7345	0.8022
10	0.4056	±0.0197	0.0635	0.5571	0.2306	±0.0218	0.0704	0.7221	0.2146	±0.0237	0.7359	0.8359
15	0.4125	±0.0230	0.0743	0.5406	0.2326	±0.0240	0.0775	0.7194	0.2097	±0.0196	0.7565	0.8247
20	0.3792	±0.0323	0.1042	0.5695	0.2243	±0.0231	0.0745	0.751	0.2049	±0.0191	0.762	0.8289
25	0.3625	±0.0308	0.0995	0.5832	0.2181	±0.0191	0.0615	0.7483	0.1979	±0.0166	0.7744	0.8303
Wilcoxon												
1	0.3903	±0.0193	0.0623	0.7015	0.3903	±0.0193	0.0623	0.7015	0.3903	±0.0193	0.0623	0.7015
2	0.3729	±0.0179	0.0576	0.74	0.3701	±0.0192	0.0621	0.7331	0.3639	±0.0184	0.0593	0.74
3	0.3687	±0.0162	0.0522	0.7497	0.3458	±0.0184	0.0593	0.7662	0.3403	±0.0211	0.068	0.7744
4	0.3701	±0.0156	0.0505	0.7455	0.3333	±0.0204	0.066	0.7799	0.3361	±0.0221	0.0714	0.7675
5	0.3708	±0.0166	0.0536	0.7359	0.3285	±0.0219	0.0706	0.7744	0.3313	±0.0251	0.0809	0.762
10	0.3604	±0.0182	0.0588	0.7428	0.3264	±0.0251	0.081	0.7648	0.3146	±0.0274	0.0884	0.7607
15	0.3486	±0.0164	0.053	0.7675	0.316	±0.0253	0.0817	0.7675	0.2993	±0.0258	0.0833	0.7717
20	0.3472	±0.0189	0.061	0.7675	0.3118	±0.0252	0.0814	0.7689	0.2937	±0.0256	0.0827	0.7758
25	0.334	±0.0218	0.0705	0.7785	0.3104	±0.0251	0.081	0.7565	0.2875	±0.0243	0.0785	0.773

a All statistics listed are averages obtained over 40 standardised train/test splits. The split ratio of 70/30 was applied to all splits.

95% CI = 95% confidence intervals on test error; **sens** = sensitivity; **spec** = specificity; **test** = test error.

graph in figure 7a shows test errors only. Figure 7b and table I (bottom half, labelled Wilcoxon) illustrate the effect of de-correlations in combination with another feature scoring criterion: the Wilcoxon rank-sum test.

The results on two univariate scores clearly illustrate the benefits of de-correlation filtering. Elimination of highly correlated features is able to boost the performance of the classifier defined on the remaining features. Stricter threshold on the MAC appears to yield better results for SAM. However, the tighter threshold does not always imply better classification accuracy. This can be observed on classifiers built for Wilcoxon scores, where MAC = 0.8 appears to improve over MAC = 0.6 when using 25 features, and we have also observed such a variability on other proteomic datasets.

Table I and figure 7 illustrate another important feature: the sensitivity of the classifier accuracy on the choice of the univariate scoring. We have tested multiple differential scoring methods (many come from microarray studies) including the Fisher-like score,[51] AUC,[15] t-test score[52,58] and simple and weighted separability scores on multiple proteomic datasets. We did not identify a clear winner, but the SAM score appears to perform very well and is consistently among the top-scoring methods.

Multivariate Greedy Features with De-Correlation Filtering

Enforcing MAC thresholds tends to improve the quality of the feature set, but the best MAC value varies from dataset to dataset and it is also different for different univariate criteria. This prompts the development of methods for choosing the MAC threshold candidate. To address this problem, instead of searching for the best MAC we have developed a new multivariate feature selection procedure that combines the advantages of univariate feature scoring and de-correlation. We refer to the new multivariate feature selection procedure as the parallel MAC method.

The parallel MAC procedure first rank-orders features using a given univariate differential expression score and then builds a feature set incrementally by choosing the best new feature from among multiple candidate features, each of them being the highest univariate score candidate at some fixed MAC level. The best feature is determined using an internal cross-validation approach. We use 10-fold cross-validation as the default. Note that the approach is different from the *classic greedy wrapper* approach that must scan and evaluate all (~60 000) possible candidate features.[59] In contrast to this, our model scans and evaluates only feature candidates that correspond to the highest ranked candidates at different MAC levels, and the number of candidates compared depends on the number of MAC levels tracked. The resolution of the method may be controlled by increasing or decreasing the number of MAC thresholds.

Table II illustrates the results of application of the parallel MAC method on the UPCI pancreatic cancer dataset. Figure 8 compares the test errors of the parallel MAC method with the fixed MAC methods for the SAM and Wilcoxon scores.

In terms of the multivariate feature selection, the parallel MAC method is a heuristic. In every step the next top feature is picked greedily from among a small number of candidates determined by the thresholds. The number of candidates can be controlled by the resolution of the MAC thresholds, thus it is independent of the total number of features.

Peak-Centred Analysis

Problems of high-dimensional profile data can be partly resolved by focusing on signal peaks and features associated with the peaks. Peak selection or peak detection methods are quite popular among SELDI-TOF-MS researchers (see Adam et al.,[13] Yasui et al.[17] and Coombes et al.[60]) in particular, because peaks, as observed in profiles, are believed to represent discrete proteins or their fragments. However, peak signatures in profiles are not perfect: individual peaks are spread over a wider area, thus the signatures of more than one peak become convolved. Low-resolution profiles can make the problem even worse: the readings recorded are already mixtures of peak signatures. In addition, peak positions may be shifted because of mass inaccuracy in different samples. As a result, exact peak positions and their intensities are hard to pinpoint. Because of these factors, the identification of peaks and definition of appropriate peak-descriptive statistics pose a challenging research problem.

In this article, we study a relatively simple peak-centred feature selection strategy that attempts to identify peak positions for sets of profiles and restricts the subsequent interpretive analysis to only such positions. The key idea we adopt is based on profile averaging: the peak position is identified on a 'mean' profile that is obtained by averaging all profiles in the training set.[4] The advantage of using the mean profile is that one can often benefit from the resulting noise reduction, which leads to a cleaner profile and more reliable peak-position estimates. The same profile averaging approach has also been recently adopted by Coombes et al.[60]

We have implemented this peak-selection procedure and tested it on the pancreatic cancer dataset. Since the number of peak positions identified in SELDI-TOF-MS profiles remains relatively large, peak selection should be combined with other feature reduc-

4 An alternative approach is to split profiles into two groups, case and control, and average them separately. This would eliminate a chance of peak cancellation. However, in this case a peak alignment procedure is necessary to merge two sets of peak positions.

Table II. Classification statistics for the support vector machine classifier combined with parallel maximum allowed correlation (MAC) feature selection (SAM [significance analysis of microarrays], and Wilcoxon rank-sum test). Results are also given for principal component analysis for comparison

No. of features	Method										
	parallel MAC threshold[a] with SAM scoring criterion				parallel MAC threshold with Wilcoxon rank-sum test				principal component analysis		
	test	95% CI	sens	spec	test	95% CI	sens	spec	test	95% CI	sens	spec
1	0.3903	±0.0199	0.0641	0.5736	0.3903	±0.0193	0.0623	0.7015	0.5179	±0.0187	0.0604	0.4566
2	0.2868	±0.0268	0.0863	0.6575	0.3535	±0.0190	0.0613	0.7634	0.4814	±0.0219	0.0707	0.5084
3	0.2479	±0.0235	0.0758	0.707	0.3472	±0.0214	0.0692	0.762	0.4457	±0.0213	0.0687	0.5406
4	0.2444	±0.0232	0.0747	0.7166	0.3493	±0.0206	0.0665	0.7524	0.4386	±0.0260	0.084	0.5616
5	0.2472	±0.0220	0.0711	0.7043	0.3431	±0.0211	0.0682	0.7607	0.4286	±0.0263	0.0849	0.5672
10	0.2118	±0.0198	0.0637	0.74	0.3264	±0.0230	0.0744	0.762	0.2607	±0.0216	0.0697	0.7339
15	0.1972	±0.0196	0.0632	0.7689	0.3069	±0.0231	0.0747	0.7813	0.2364	±0.0245	0.079	0.7647
20	0.2	±0.0191	0.0617	0.7717	0.2944	±0.0236	0.0762	0.7882	0.2171	±0.0176	0.0567	0.7941
25	0.2035	±0.0193	0.0624	0.762	0.2917	±0.0252	0.0813	0.7854	0.2129	±0.0191	0.0617	0.7997

a The method applied used 13 MAC thresholds dividing the interval [0.4, 1] into equal size components in increments of 0.05.

95% CI = 95% confidence intervals on test error; **sens** = sensitivity; **spec** = specificity; **test** = test error.

tion strategies. Figure 9 shows the positions of the highest-ranked 15 peaks identified by the parallel MAC method restricted to peaks on the mean case and control profiles.

Table III shows the performance of SVM classifiers restricted to peaks selected by the SAM and Wilcoxon differential feature selection measurements. Comparing these with whole-profile analysis results in figure 8, we see that test errors on the SVM have improved in one case but are comparable in the other. However, we have seen cancer datasets where the whole-profile analysis outperformed the peak-only analysis, so, in general, we cannot say that one is always preferred to the other. The reasons why the whole-profile analysis may become better is that the peaks in the low-resolution SELDI-TOF-MS profiles may overlap and the discriminative information can be hidden anywhere (in valleys, slopes of the profile, etc.). Thus, pure focus on peaks carries a threat of information loss: some important discriminative clues in profiles may be overlooked. To be thorough, both analysis modes should be explored.

As can be seen from the differences in figure 8 and figure 10, de-correlation can help us achieve a lower error in both modes of analysis. Benefits in whole-profile analysis are more obvious, when correlations between covariates can be frequent. In the case of peaks, double- and triple-charged ions in the signal may represent full correlations between peaks, and restriction to peak positions alone may result in redundant information. By applying the de-correlation filter to peak-selected covariates, we can eliminate these uninformative peaks, which further reduces our search space for features.

Multivariate Features Based on Principal Component Analysis

An alternative approach to multivariate feature selection is offered by multivariate projection techniques such as PCA[21] or ICA.[22,23] PCA is a widely used method for reducing the number of dimensions of a dataset. The PCA computes projections of a high-dimensional data into a lower dimensional subspace such that the variance retained in the projected data is maximised. Equivalently, the PCA gives uncorrelated projected distributions and minimises the least-square reconstruction error. From the proteomic profile perspective, PCA identifies orthogonal sets of correlated features and constructs composite features (components) that are uncorrelated but tend to explain most of the observed variance in the data.

Figure 11a shows the result of the SVM classifier trained with up to 25 PCA features that correspond to the eigenvectors of the data matrix with the highest eigenvalues. Likewise, figure 11b displays the same results on peak-selected data. The principal

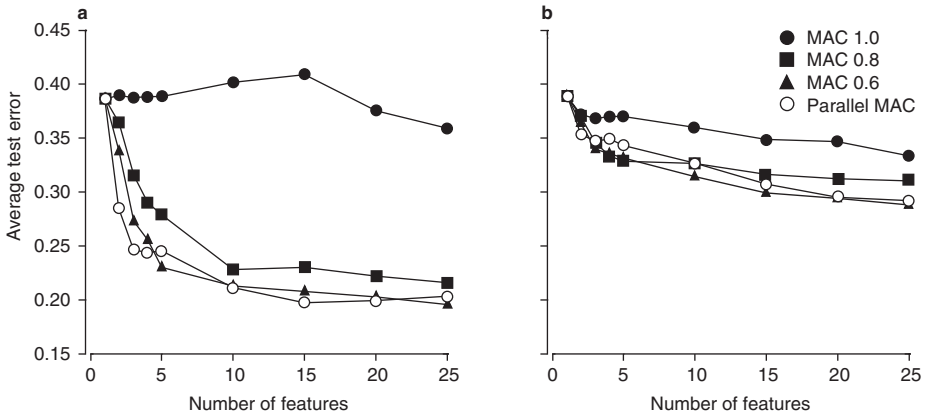

Fig. 8. A comparison of the parallel maximum allowed correlation (MAC) method to classifiers with MAC = 0.6, 0.8 and 1.0. (**a**) Feature selection with the SAM (significance analysis of microarrays) score. (**b**) Feature selection with the Wilcoxon rank-sum score. Test errors are given for a varied number of features and for different de-correlation (MAC) thresholds. Additional performance statistics are given in table I and table II.

component projections enable classification of case versus control with test errors comparable with the multivariate method with parallel MAC thresholds.

The limitation of the PCA approach is the interpretation of the features. Identification of profile positions responsible for good discriminatory performance is much more difficult than the earlier techniques (univariate differential expression) that let us select k profile positions (biomarkers) directly. The PCA components tell us what positions are more important for the first, second, etc. projections, but each of these is a combination of features that are on equal or very similar levels.

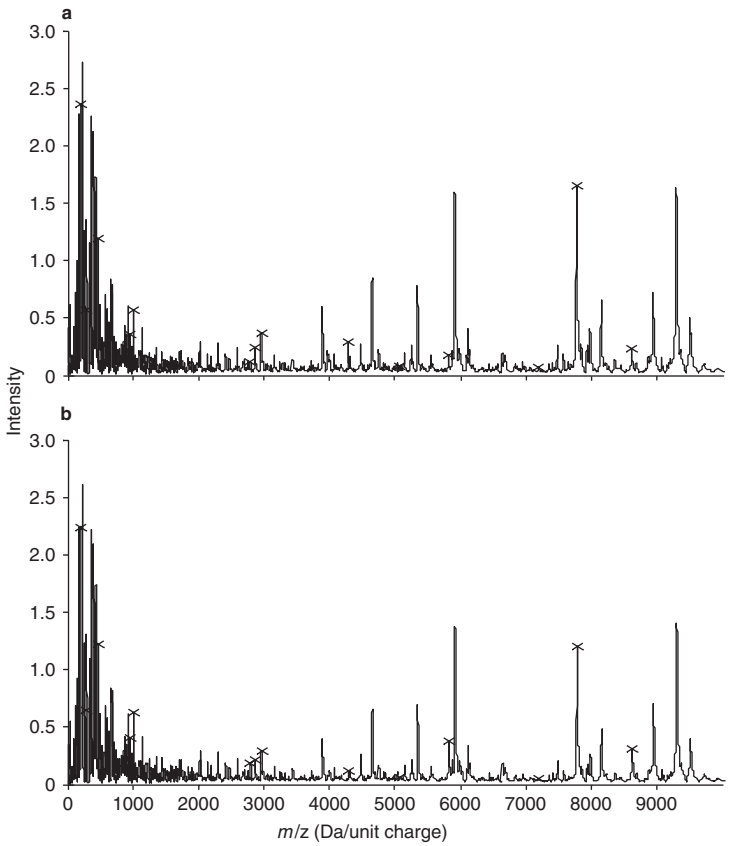

Fig. 9. Positions of the 15 highest-ranking SAM (significance analysis of microarrays) features (crosses on graphs) restricted to peaks: (**a**) mean control profile and (**b**) mean case profile. The positions available for feature selection are limited to local maxima. ***m*/*z*** = mass-to-charge ratio.

Table III. Classification statistics for univariate feature scoring methods (SAM [significance analysis of microarrays], and Wilcoxon rank-sum test) with de-correlations and the linear support vector machine classifier on peak-only analysis. Three maximum allowed correlation (MAC) thresholds (0.6, 0.8 and 1.0) are compared for a varied number of features (1–25)[a]

Method used and no. of features	MAC = 1.0				MAC = 0.8				MAC = 0.6			
	test	95% CI	sens	spec	test	95% CI	sens	spec	test	95% CI	sens	spec
SAM												
1	0.3903	±0.0199	0.0641	0.5736	0.3903	±0.0199	0.0641	0.5736	0.3903	±0.0199	0.0641	0.5736
2	0.3653	±0.0260	0.0838	0.5997	0.366	±0.0272	0.0878	0.5818	0.3188	±0.0333	0.1073	0.6327
3	0.3361	±0.0276	0.0889	0.6272	0.3	±0.0314	0.1015	0.6602	0.2458	±0.0274	0.0884	0.7153
4	0.3188	±0.0299	0.0964	0.6451	0.2458	±0.0268	0.0864	0.7249	0.2319	±0.0266	0.0857	0.7387
5	0.2972	±0.0304	0.0981	0.663	0.2326	±0.0253	0.0817	0.729	0.2153	±0.0212	0.0683	0.7552
10	0.209	±0.0276	0.0892	0.7662	0.2083	±0.0211	0.068	0.7552	0.1924	±0.0198	0.064	0.7744
15	0.1972	±0.0218	0.0703	0.7772	0.1868	±0.0186	0.06	0.7895	0.1903	±0.0182	0.0589	0.7785
20	0.1875	±0.0200	0.0644	0.7882	0.1854	±0.0164	0.0528	0.7868	0.1854	±0.0189	0.0611	0.7895
25	0.1854	±0.0247	0.0797	0.8047	0.1833	±0.0191	0.0616	0.7923	0.1861	±0.0188	0.0607	0.7909
Wilcoxon												
1	0.3854	±0.0168	0.0542	0.718	0.3854	±0.0168	0.0542	0.718	0.3854	±0.0168	0.0542	0.718
2	0.3722	±0.0220	0.0708	0.7387	0.3472	±0.0226	0.0728	0.762	0.3375	±0.0207	0.0667	0.7758
3	0.3597	±0.0228	0.0736	0.7524	0.3312	±0.0224	0.0724	0.7758	0.3083	±0.0271	0.0873	0.7428
4	0.3514	±0.0200	0.0646	0.7524	0.316	±0.0236	0.0762	0.7552	0.2944	±0.0290	0.0937	0.7428
5	0.3465	±0.0310	0.1001	0.7331	0.3146	±0.0265	0.0856	0.7524	0.2826	±0.0299	0.0966	0.7428
10	0.3222	±0.0280	0.0902	0.7331	0.2986	±0.0306	0.0988	0.74	0.2472	±0.0298	0.0962	0.7675
15	0.3236	±0.0289	0.0932	0.7249	0.2937	±0.0317	0.1022	0.74	0.2257	±0.0272	0.0879	0.7662
20	0.2972	±0.0309	0.0997	0.729	0.2847	±0.0324	0.1046	0.751	0.2104	±0.0238	0.0768	0.7675
25	0.2896	±0.0303	0.0978	0.7304	0.2687	±0.0315	0.1017	0.7675	0.2007	±0.0206	0.0664	0.7662

a All statistics listed are averages obtained over 40 standardised train/test splits. The split ratio of 70/30 was applied to all splits.

95% CI = 95% confidence intervals on test error; **sens** = sensitivity; **spec** = specificity; **test** = test error.

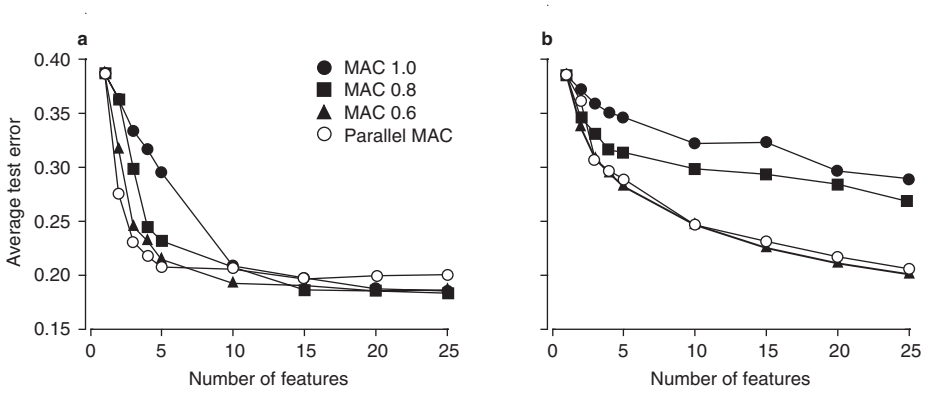

Fig. 10. A comparison of the parallel maximum allowed correlation (MAC) method to classifiers with MAC = 0.6, 0.8 and 1.0 when restricted to peaks. (**a**) Feature selection with the SAM (significance analysis of microarrays) score. (**b**) Feature selection with the Wilcoxon rank-sum score. Test errors are given for varied number of features and for different de-correlation (MAC) thresholds. The corresponding data are given in table III and table IV.

Permutation-Based Validation

The results presented above show that it is possible to learn predictive models that can achieve a low classification error on the SELDI-TOF-MS samples. To support the significance of these results, particularly the fact that the sample profiles carry useful discriminative signal, one may seek 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 data labelling. Figure 12 shows the result of the random permutation test for the parallel MAC classifiers analysed in figure 8 and figure 10. The figure plots the estimate of the mean test error one would obtain by learning the classifier on 5–25 features for randomly assigned class labels, and estimates of 95% and 99% test error bounds. The estimates are obtained using 100 random class-label permutations of the original pancreatic cancer 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. However, we note that the permutation test does not protect against data biases and possible confounding introduced through case/control subject selection, sample collection or sample preprocessing steps. A more thorough exploration of permutation-based validation methods can be found in Lyons-Weiler et al.[45]

Conclusions

In a recently published discussion, Diamandis[61] and Petricoin and Liotta[62] shared their views on the likely research and clinical utility of SELDI-TOF-MS profiling. Diamandis[61] raises a number of concerns about the utility of SELDI-TOF-MS technology. The work in this article addresses some of these concerns. First,

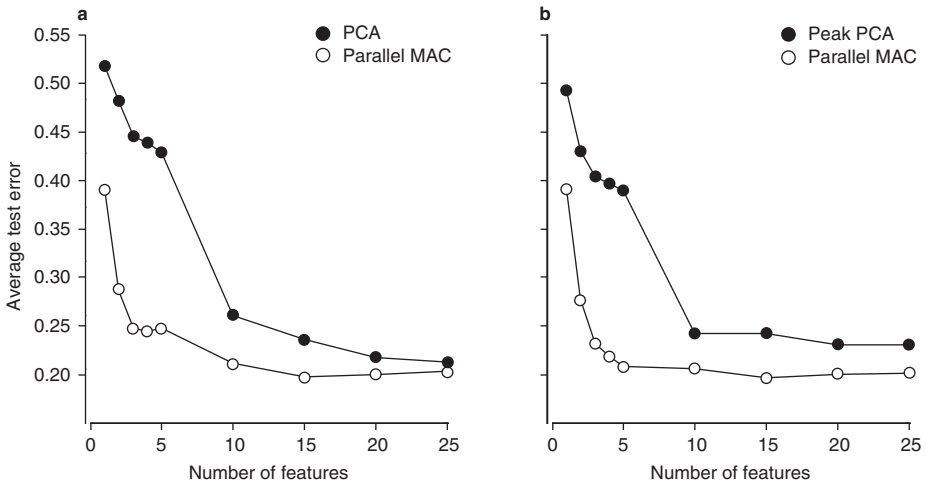

Fig. 11. Comparison of discriminability of features obtained by the parallel maximum allowed correlation (MAC) and principal component analysis (PCA) approaches: (**a**) whole-profile analysis and (**b**) peak-only analysis. The performance of this technique is plotted against the performance of the SAM (significance analysis of microarrays) score classifier for the purpose of comparison. The corresponding data are given in tables I to IV.

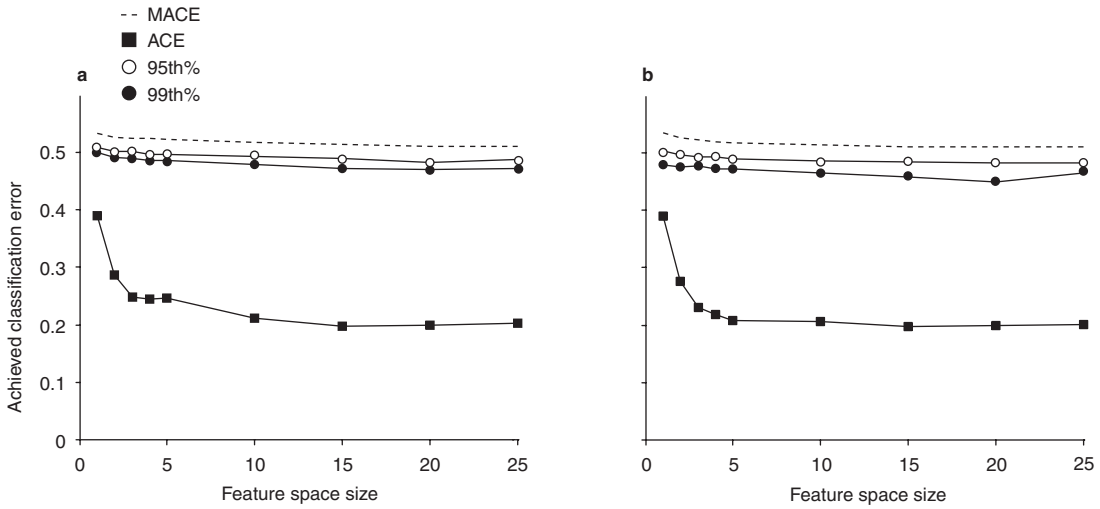

Fig. 12. An example of the permutation-based analysis for the parallel maximum allowed correlation approach with SAM (significance analysis of microarrays) features in the (**a**) whole-profile mode and (**b**) peak-only mode. The dashed line indicates mean classification errors (MACE) for the model under the null hypothesis: the class labels in the data are assigned to profiles randomly. The black and white circles indicate the estimate of the 95th percentile and the 99th percentile, respectively, of the test error statistic under the null hypothesis. The squares indicate achieved classification errors (ACE) using the original data labelling.

the results show that SELDI-TOF-MS-based profiling can indeed lead to good predictive case-control models. Second, it shows that feature selection that takes into account characteristics of the spectra, particularly correlations among intensity measurements in the profile, helps to improve the discriminatory performance. The presence of correlates in the signal and their substitutability also explains the variety of possible *m*/*z* feature panels with comparable discriminatory power.

Multivariate feature selection is the key to our ability to learn the discriminative patterns in cancer datasets available today. In this study we have focused on statistical techniques to efficiently identify a small set of highly discriminative features. We have demonstrated a good discriminative performance of a number of relatively simple feature selection methods on the UPCI pancreatic cancer dataset. Especially useful is our de-correlation enhancement of the univariate feature selection methods. The de-correlation approach tends to improve the performance of feature selection methods in both the whole-profile and peak-only modes on pancreatic cancer data. We note that we have obtained and observed similar improvements on other UPCI and publicly available datasets, including lung, melanoma, prostate and ovarian cancer datasets. It is very likely that the introduction of additional knowledge, such as peptide identification of *m*/*z* values, can further improve the performance of feature selection methods.

Many research issues related to SELDI-TOF-MS datasets remain open. More work is needed to optimise data preprocessing steps, particularly profile rescaling. Thorough evaluation of benefits of all preprocessing steps and their order in the removal of systematic sources of errors is needed. Very little is presently known regarding possible non-instrument sources of variations (data collection, data storage, surface irregularities). Understanding these sources may lead to a better experimental design and improved ability to detect the true discriminative information. Finally, the main focus of this study was on analysis of multivariate feature selection methods. To keep it simple we restricted our analysis to only linear SVM classifiers, which nevertheless may have affected the accuracy of profile classifications. Although preliminary results on the pancreatic cancer dataset indicate that the technology is able to discriminate well between the cases and controls, higher sensitivities and specificities are needed to make the SELDI-TOF-MS screening a clinically useful approach. The evaluation of a diversity of complex classification models in combination with multivariate feature selection methods is needed to assess the true potential and detection limits of the SELDI technology on these data.

The SELDI-TOF-MS technology of Ciphergen Biosystems Inc. yields relatively low-resolution profiles, which are likely to limit its applicability at some point. For example, a limited instrument resolution may not be able to detect higher resolution signals, and potentially useful differences therein may remain hidden and convolved with other signals. Increasing use of new high mass resolution, matrix-assisted laser desorption/ionisation (MALDI)/SELDI-TOF-MS instruments will provide larger, and more analytically challenging, datasets to test these questions. In addition, a low-resolution profile is harder to use for the protein identification task. Thus, higher resolution instruments, with integrated peak

Table IV. Classification statistics for the support vector machine classifier combined with parallel maximum allowed correlation (MAC) feature selection (SAM [significance analysis of microarrays], and Wilcoxon rank-sum test) on peak-only analysis. Results are also given for principal component analysis for comparison

No. of features	Method										
	parallel MAC threshold[a] with SAM scoring criterion				parallel MAC threshold with Wilcoxon rank-sum test				principal component analysis		
	test	95% CI	sens	spec	test	95% CI	sens	spec	test	95% CI	sens	spec
1	0.3903	±0.0199	0.0641	0.5736	0.3854	±0.0168	0.0542	0.718	0.4921	±0.0217	0.0701	0.5014
2	0.2764	±0.0278	0.0898	0.663	0.3625	±0.0258	0.0832	0.7276	0.4307	±0.0166	0.0535	0.5532
3	0.2313	±0.0203	0.0655	0.7139	0.3069	±0.0312	0.1008	0.7428	0.4029	±0.0237	0.0765	0.5602
4	0.2181	±0.0229	0.074	0.7345	0.2972	±0.0322	0.1038	0.7538	0.3957	±0.0199	0.0642	0.5756
5	0.2083	±0.0236	0.0763	0.7359	0.2896	±0.0337	0.1088	0.7538	0.3886	±0.0218	0.0702	0.6064
10	0.2056	±0.0229	0.0738	0.7552	0.2472	±0.0328	0.106	0.773	0.2414	±0.0211	0.0682	0.7255
15	0.1965	±0.0199	0.0643	0.7593	0.2312	±0.0293	0.0946	0.7675	0.2421	±0.0185	0.0596	0.7283
20	0.1993	±0.0220	0.0709	0.7675	0.2167	±0.0222	0.0715	0.7744	0.23	±0.0192	0.062	0.7507
25	0.2007	±0.0234	0.0754	0.7675	0.2062	±0.0200	0.0644	0.7909	0.2307	±0.0195	0.0629	0.7465

a The method applied used 13 MAC thresholds dividing the interval [0.4, 1] into equal size components in increments of 0.05.

95% CI = 95% confidence intervals on test error; **sens** = sensitivity; **spec** = specificity; **test** = test error.

identification capability, are needed (and are increasingly available) for true biomarker discovery. However, we believe that the understanding of the nature of a proteomic signal, possible sources of variations and techniques for their removal, and multivariate data analysis methods developed in this work will also apply to other profiling platforms.

Acknowledgements

This work was supported in part by the Early Detection Research Network grant no. UO1 CA84968 to William L. Bigbee, and Lung SPORE (Specialized Programs of Research Excellence) grant no. P50 CA90440 to Jill M. Siegfried (which supported some members of our team).

The authors have no conflicts of interest that are directly relevant to the content of this article.

References

1. Petricoin EF, Ardekani AM, Hitt BA, et al. Use of proteomic patterns in serum to identify ovarian cancer. Lancet 2002; 359: 572-7
2. Wright Jr GW, Cazares LH, Leung SM, et al. Proteinchip(R) surface enhanced laser desorption/ionization (SELDI) mass spectrometry: a novel protein biochip technology for detection of prostate cancer biomarkers in complex protein mixtures. Prostate Cancer Prostatic Dis 1999; 2 (5/6): 264-76
3. Adam BL, Vlahou A, Semmes OJ, et al. Proteomic approaches to biomarker discovery in prostate and bladder cancers. Proteomics 2001; 1: 1264-70
4. Zhu W, Wang X, Ma Y, et al. Detection of cancer-specific markers amid massive mass spectral data. Proc Natl Acad Sci U S A 2003; 100: 14666-71
5. Jones MB, Krutzsch H, Shu H, et al. Proteomic analysis and identification of new biomarkers and therapeutic targets for invasive ovarian cancer. Proteomics 2002; 2: 76-84
6. Li J, Zhang Z, Rosenzweig J, et al. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. Clin Chem 2002; 48: 1296-304
7. Watkins B, Szaro R, Ball S, et al. Detection of early stage cancer by serum protein analysis. Am Lab 2001; 6: 32-6
8. Wadsworth JT, Somers K, Stack B, et al. Identification of patients with head and neck cancer using serum protein profiles. Arch Otolaryngol Head Neck Surg 2004 Jan; 130: 98-104
9. Poon TC, Yip TT, Chan AT, et al. Comprehensive proteomic profiling identifies serum proteomic signatures for detection of hepatocellular carcinoma and its subtypes. Clin Chem 2003 May; 49 (5): 752-60
10. Zhukov TA, Johanson RA, Cantor AB, et al. Discovery of distinct protein profiles specific for lung tumors and pre-malignant lung lesions by SELDI mass spectrometry. Lung Cancer 2003; 40: 267-79
11. Xiao XY, Tang Y, Wei XP, et al. A preliminary analysis of non-small cell lung cancer biomarkers in serum. Biomed Environ Sci 2003; 16: 140-8
12. Kozak KR, Amneus MW, Pusey SM, et al. Identification of biomarkers for ovarian cancer using strong anion-exchange ProteinChips: potential use in diagnosis and prognosis. Proc Natl Acad Sci U S A 2003 Oct; 100 (21): 12343-8
13. Adam BL, Qu Y, Davis JW, et al. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. Cancer Res 2002; 62: 3609-14
14. Petricoin E, Ornstein DK. Serum proteomic patterns for detection of prostate cancer. J Natl Cancer Inst 2002; 94 (20): 1576-8
15. Qu Y, Adam BL, Yasui Y, et al. Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. Clin Chem 2002; 48: 1835-43
16. Qu Y, Adam B, Thornquist M, et al. Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensionality data. Biometrics 2003; 59: 143-51

17. Yasui Y, Pepe M, Thompson ML, et al. A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. Biostatistics 2003; 4: 449-63
18. Carpenter M, Melath S, Zhang S, et al. Statistical processing and analysis of proteomic and genomic data [online]. Available from URL: http://www.pharmasug.org/2003/BestPapers/sp106.pdf [Accessed 2004 Jan]
19. Sidransky D, Irizarry R, Califano JA, et al. Serum protein MALDI profiling to distinguish upper aerodigestive tract cancer patients from control subjects. J Natl Cancer Inst 2003; 95: 1711-7
20. Jain AK, Dubes RC. Algorithms for clustering data. Englewood Cliffs (NJ): Prentice-Hall, 1988
21. Jolliffe IT. Principal component analysis. New York: Springer-Verlag, 1986
22. Jutten C, Herault J. Blind separation of sources 1: an adaptive algorithm based on neuromimetic architecture. Signal Process 1991; 24 (1): 1-10
23. Lee TW. Independent component analysis: theory and applications. Boston (MA): Kluwer Academic Publishers, 1998
24. Grizzle WE, Adam BL, Bigbee WL, et al. Serum protein expression profiling for cancer detection: validation of a SELDI-based approach for prostate cancer. Dis Markers 2004; 19: 185-95
25. Baggerly KA, Morris JS, Coombes KR. Reproducibility of SELDI-TOF protein patterns in serum: comparing data sets from different experiments. Bioinformatics 2004; 20 (5): 777-85
26. Durbin BP, Hardin JS, Hawkins DM, et al. A variance-stabilizing transformation for gene-expression microarray data. Bioinformatics 2002; 18: 105-10
27. Sankoff D, Kruskal J. Time warps, string edits, and macromolecules: the theory and practice of sequence comparison. Reading (MA): Addison-Wesley, 1983
28. Sakoe H, Chiba S. Dynamic programming optimization for spoken word recognition. IEEE Trans Acoust 1978 Feb; 26: 43-9
29. Eilers PHC. Parametric time warping. Anal Chem 2004; 76: 404-11
30. Ramsey JO, Li X. Curve registration. J R Stat Soc Ser B 1998; 60: 351-63
31. Semmes OJ, Feng Z, Adam BL, et al. Evaluation of serum protein profiling by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry for the detection of prostate cancer: I. Assessment of platform reproducibility. Clin Chem 2005 Jan; 51 (1): 102-12
32. Grizzle WE, Meleth S, Eltoum IA, et al. Novel approaches to smoothing and comparing SELDI TOF spectra. Cancer Inform 2004; 1 (1): 78-85
33. Semmes OJ, Feng Z, Adam B-L, et al. Evaluation of serum protein profiling by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry for the detection of prostate cancer: I. Assessment of platform reproducibility. Clin Chem 2005; 51: 102-12
34. Breiman L, Friedman JH, Olshen RA, et al. Classification and regression trees. Belmont (CA): Wadsworth, 1984
35. Quinlan JR. C4.5: programs for machine learning. San Francisco (CA): Morgan Kaufmann, 1993
36. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York: Springer, 2001
37. Burgess C. A tutorial on support vector machines for pattern recognition. Mach Learn J 1998; 2: 121-67
38. Vapnik VN. The nature of statistical learning theory. New York: Springer-Verlag, 1995
39. Scholkopf B, Smola A. Learning with kernels. Boston (MA): MIT Press, 2002
40. Bayes T. An essay towards solving a problem in the doctrine of chances. Philos Trans R Soc Lond 1763; 53: 370-418
41. Russel S, Norvig P. Artificial intelligence: a modern approach. Englewood Cliffs (NJ): Prentice Hall, 2002
42. Ball G, Mian S, Holding F, et al. An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumors and rapid identification of potential biomarkers. Bioinformatics 2002; 18 (3): 395-404
43. Haykin S. Neural networks. New York: Macmillan, 1994
44. Bishop C. Neural networks for pattern recognition. Oxford: Oxford University Press, 1995
45. Lyons-Weiler J, Pelikan R, Zeh III HJ, et al. Assessing the statistical significance of the achieved classification error of classifiers constructed using serum peptide profiles and a prescription for random resampling repeated studies for massive high-throughput genomic and proteomic studies. Cancer Inform 2005; 1 (1): 53-77
46. Good P. Permutation tests: a practical guide to resampling methods for testing hypothesis. New York: Springer-Verlag, 1994
47. Kendall MG. The treatment of ties in ranking problems. Biometrika 1945; 33: 239-51
48. Goland P, Fischl B. Permutation tests for classification: towards statistical significance in image-based studies. The 18th International Conference on Information Processing in Medical Imaging. New York: Springer-Verlag, 2003; 2732: 330-41. Lecture Notes in Computer Science
49. Duda RO, Hart PE, Stork DG. Pattern classification. 2nd ed. New York: John Wiley and Sons, 2000
50. Mangasarian OL, Musicant DR. Lagrangian support vector machines. J Mach Learn Res 2001; 3: 161-77
51. Fisher R. The use of multiple measurements in taxonomic problems. Ann Eugen 1936; 7: 79-188
52. Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics 2001; 17 (6): 509-19
53. Hanley J, McNeil B. The meaning and use of the area under a receiver operating characteristic curve. Diagn Radiol 1982; 143 (1): 29-36
54. Cover TH, Thomas JA. Elements of information theory. New York: Wiley-Interscience, 1991
55. Bonnlander BV, Weigend AS. Selecting input variables using mutual information and nonparametric density estimation. International Symposium on Artificial Neural Networks (ISANN); Tainan, Taiwan; 1994 Dec 15-17
56. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 1995; 57: 289-300
57. Tusher V, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001; 98 (9): 5116-21
58. Student. The probable error of a mean. Biometrika 1908; 6: 1-25
59. Kohavi R, John GH. The wrapper approach. In: Liu H, Motoda H, editors. Feature selection for knowledge discovery in databases. New York: Springer-Verlag, 1998
60. Coombes KR, Tsavachidis S, Morris JS, et al. Improved peak detection and quantication of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. Houston (TX): University of Texas, 2004. MD Anderson Biostatistics Technical Report no.: UTMDABTR-001-04
61. Diamandis EP. Point: proteomic patterns in biological fluids: do they represent the future of cancer diagnostics? Clin Chem 2003; 49: 1272-5
62. Petricoin III E, Liotta LA. Counterpoint: the vision for a new diagnostic paradigm. Clin Chem 2003; 49: 1276-8

Correspondence and offprints: Dr *Milos Hauskrecht*, Department of Computer Science, 5329 Sennott Building, University of Pittsburgh, Pittsburgh, PA 15260, USA.
E-mail: milos@cs.pitt.edu