

RESEARCH ARTICLE

Annotated regions of significance of SELDI-TOF-MS spectra for detecting protein biomarkers

Chuen Seng Tan^{1,2}, Alexander Ploner¹, Andreas Quandt¹, Janne Lehtiö^{3,4},
Maria Pernemalm^{3,4}, Rolf Lewensohn^{3,4} and Yudi Pawitan¹

¹ Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

² Center for Molecular Epidemiology, Yong Loo Lin School of Medicine, National University of Singapore and Genome Institute of Singapore, Singapore

³ Cancer Centrum Karolinska, Karolinska Institutet, Stockholm, Sweden

⁴ Clinical Proteomics, Karolinska Biomics Center, Karolinska University Hospital, Stockholm, Sweden

Peak detection is a key step in the analysis of SELDI-TOF-MS spectra, but the current default method has low specificity and poor peak annotation. To improve data quality, scientists still have to validate the identified peaks visually, a tedious and time-consuming process, especially for large data sets. Hence, there is a genuine need for methods that minimize manual validation.

We have previously reported a multi-spectral signal detection method, called RS for 'region of significance', with improved specificity. Here we extend it to include a peak quantification algorithm based on annotated regions of significance (ARS). For each spectral region flagged as significant by RS, we first identify a dominant spectrum for determining the number of peaks and the m/z region of these peaks. From each m/z region of peaks, a peak template is extracted from all spectra *via* the principal component analysis. Finally, with the template, we estimate the amplitude and location of the peak in each spectrum with the least-squares method and refine the estimation of the amplitude *via* the mixture model.

We have evaluated the ARS algorithm on patient samples from a clinical study. Comparison with the standard method shows that ARS (i) inherits the superior specificity of RS, and (ii) gives more accurate peak annotations than the standard method. In conclusion, we find that ARS alleviates the main problems in the preprocessing of SELDI-TOF spectra.

The R-package ProSpect that implements ARS is freely available for academic use at <http://www.meb.ki.se/~yudipaw>.

Received: July 12, 2006

Accepted: August 12, 2006



Keywords:

Peak annotation / Peak detection / SELDI / Signal detection

Correspondence: Professor Yudi Pawitan, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, P.O. Box 281, SE-171 77 Stockholm, Sweden

E-mail: yudi.pawitan@ki.se

Fax: +46-8-31 49 57

Abbreviations: ANOVA, analysis of variance; ARS, annotated regions of significance; FDR, false discovery rate; MSE, mean squared error; RS, regions of significance; SSA, simultaneous spectrum analysis; SPF, simple peak finding

1 Introduction

Proteomic profiles from surface-enhanced laser desorption and ionization (SELDI) have been used for clinical biomarker discovery with promising results, *e.g.* for identifying patients with early stage cancer [1–3]. SELDI protein profiling was first described in 1993 [4] and has been developed subsequently by Ciphergen Biosystems. In SELDI-based profiling, retentate chromatography and mass

spectrometry are combined, allowing direct mass spectrometry based protein profiling of biological samples. The raw spectral data generated is pre-processed in several calibration and adjustment steps that yield two quantities, the intensity (ion abundance) and the mass *per charge* (m/z) registered at each time point (see [5] for the Ciphergen default method or [6] and [7] for more recent alternative suggestions). A protein profile of a sample is the collection of all pairs of intensity and m/z -value measured on the sample, with peaks in the intensity characterizing the presence of a protein or peptide.

SELDI spectra usually contain high frequency noise and consequently a multitude of misleading peaks that are entirely due to measurement error, chemical and other background noise. The next step in the SELDI work-flow is therefore usually the identification of peaks that represent genuine biological differences between samples. Once these peaks have been identified in each spectrum, the actual peak intensities still need to be determined. This latter step is referred to as peak annotation or peak labeling, and it is often complicated by slight misalignments between spectra along the m/z axis, which may be due to either subtle biological signal modifications or simple technical measurement variability.

The currently most widely-used peak detection method combines the peak identification and annotation steps into one procedure. It has been introduced by Ciphergen and implemented in their ProteinChip software. This algorithm, which will be referred to as the standard or default method, has an unfortunate tendency to find too many false peaks (*i.e.*, low specificity) and to mislabel even some of the real peaks it detects. By using the results of the standard method uncritically, scientists risk both missing relevant peaks and including biologically meaningless noise into the downstream biomarker discovery process. In practice, researchers usually try to verify that the identified peaks are real and correctly annotated by visually going through each of the spectra in parallel. Much of this time-consuming work can be avoided, if specificity and peak labeling can be sufficiently improved.

Dissatisfaction with the standard method has led to numerous proposed alternatives (*e.g.* [8–11], see also Section 4.2). We have previously reported a peak detection method that uses a multi-spectral approach for identifying regions containing potential biomarkers [12]; we have demonstrated that this signal detection algorithm, called RS for ‘region of significance’, has better operating characteristics than the standard and several other methods. In contrast with many existing methods, RS takes advantage of the information from all spectra simultaneously and gives the user an objective control of the false discovery rate (FDR) among the reported regions.

In this paper, we describe a new annotation algorithm, referred to as annotated regions of significance (ARS), that labels peaks in regions previously identified as significant by RS. This reduces the risk of trying to identify peaks where

there is little or no biological signal, which is the most common reason for spurious peaks. Furthermore, by focusing on regions with strong signal, this approach allows us to identify peak templates from all spectra without assuming a specific parametric shape through the principal component analysis (PCA) [13]. A similar procedure was applied in the context of nuclear magnetic resonance spectral data [14]. All spectra are then fitted to the template through the least-squares method that estimates the location and amplitude of the peaks. In addition, the mixture model is applied to the estimation of amplitude because it allows for differences in peak shapes among spectra but all coming from the same peptide/protein.

We have applied our new method to SELDI profiles of serum collected from lung cancer patients. Comparison with the standard method shows that our novel approach has better sensitivity and is more accurate in labeling peaks. The complete ARS algorithm is implemented in the package *ProSpect* for the statistical computing language R. The package performs both the signal detection and peak quantification steps, offers graphical displays of the results, and allows the export of the detected regions of peaks to the ProteinChip software. The package is freely available for academic use at <http://www.meb.ki.se/~yudpaw>.

2 Materials and Methods

2.1 Tissue samples, sample preparation and chemicals

Serum samples were obtained from lung cancer patients (stadium IIIa). The study was approved by the local ethical research committee, and samples were obtained with the patients’ consent. The sera were prepared according to standard protocol and stored at -80°C until analysis. The study included eight serum samples from patients diagnosed with adenocarcinoma and eight serum samples from patients diagnosed with squamous-cell carcinoma.

2.2 SELDI analysis of serum samples with standard method

Serum was fractionated prior to protein profiling to reduce sample complexity. Weak Cationic Exchange (CM10) ProteinChip Arrays (Ciphergen, CA, USA) were used throughout the study. Fractionated serum was diluted 1:5 in binding buffer containing 50 mM ammonium acetate pH 4.5, 0.1% w/v Triton X-100. Duplicates of 100 μL of diluted sample were applied to CM10 chips and incubated at 4°C for 1.5 hours. The ProteinChip surface was washed twice with the appropriate buffer, briefly rinsed with water and air-dried. Two times 1 μL of 20% saturated alpha-cyano-4-hydroxy cinnamic acid in 50% acetonitrile and 0.5% trifluoro-acetic acid were applied to each spot. ProteinChip arrays were analysed using a ProteinChip analyser

(TOF MS), Protein Biology System IIc (Ciphergen, CA, USA). Data was collected using the automated chip protocol. For detection of low molecular weight proteins, the protocol settings were: high mass 100 kDa with optimized mass between 2–10 kDa, laser intensity 127, detector sensitivity 8; 70 shots were averaged for each spot.

Standard ProteinChip software data analysis was done following baseline subtraction (using 8-times expected peak-width), peak intensities were normalized using total-ion-current (using external normalization coefficient 1.0). Biomarker wizard was used for peak detection using the following settings: first-pass detection S/N 5, present in at least 5% of spectra, mass difference of 0.3% allowed and second-pass detection at S/N 2. Altogether 89 peak regions were detected. ‘Min Peak Threshold’ was set to a low level because we expected more heterogeneity in spectra from patient samples than from e.g. cell lines. The setting for the ‘first pass’ was obtained by an experienced SELDI analyst (MP), by visually inspecting the spectra at various settings and choosing one that minimized false positives.

2.3 Peak Detection Algorithm

2.3.1 Signal Detection Algorithm: RS

We first describe briefly the method we use in [12] to identify the regions of significant variation between spectra. Baseline subtraction and normalization are applied as described above. The signal detection algorithm performs some additional pre-processing, specifically alignment of the m/z values and an additional robust baseline correction. In the next step, it constructs windows that divide the spectra along the m/z axis in a consecutive and non-overlapping fashion, with the same number of measurements in each window. The algorithm then computes a F^* -statistic for each window. This F^* is a modification of the standard F -statistic from a one-way analysis of variance (ANOVA) model with spectra as factor. F^* is designed to test for intensity differences between spectra irrespective of the underlying biology of the samples. The modifications made to the F -statistic increase its robustness and adjust for the dependency between closely neighboring measurements. The distribution of F^* has been verified empirically, using blank chips without biological material. To account for multiple testing, instead of using traditional p -values, the significance of differences between spectra is assessed via the false discovery rate (FDR), defined as the expected proportion of false positives among the significant results.

The signal detection step outputs a FDR value for each window. For further processing, we select only windows with FDR less than a pre-specified cutoff value (usually 5%). Selected windows that are contiguous along the m/z axis are merged, and the resulting distinct spectral regions are called clusters. To illustrate this, for the clinical data described in Section 2.1, we started with an initial set of 1053 windows with five measurements each; 302 of these were found to

have FDR below 5%; after merging contiguous windows we obtained 102 distinct clusters containing between 5 and 55 measurements.

2.3.2 Peak Quantification Algorithm: ARS

The proposed peak quantification algorithm for peak annotation is run separately on each of the clusters identified during the signal detection step. The algorithm comprises two distinct steps: (i) identifying a template across all spectra in each m/z range of peak *via* PCA [13], and (ii) fitting the template to the spectra *via* the least-squares method and mixture model.

We have investigated a simple alternative procedure for Step 1 based on choosing the spectrum with the strongest signal as the template; see the Supplementary Material for details. The PCA-based algorithm is less *ad hoc* and better in the sense that it automatically chooses the best fitting template from among the collection of spectra in a cluster. In Step 2, we estimate the amplitude and m/z location of the peak *via* least-squares method with a refinement made to the estimation of amplitude *via* the mixture model. We illustrate the detailed application of the individual steps for a specific cluster from the clinical data in the Supplementary Material.

The cluster for analysis, consisting of contiguous windows constructed in RS, is not guarantee to cover an entire peak shape. To overcome this, before applying the peak quantification algorithm, we extend the spectra appropriately along the m/z dimension. To simplify this is done on the spectrum with a strong signal (dominant spectrum) in the cluster. After the extension, we identify m/z region of each peak by using a modification of the simple peak finding (SPF) algorithm described in [9] to the dominant spectrum. Details of this algorithm are in the Supplementary Material.

Step 1: Identification of a template for each peak region

We perform PCA across all spectra in the m/z region of each peak to obtain a template that best captures the peak shape. Assuming the spectra $S(t)$'s have a common template $f(t)$ with varying amplitudes A 's, but are misaligned due to shifts δ 's, we can express each spectrum as

$$S(t) = \beta_0 + Af(t - \delta). \quad (1)$$

The β_0 is an additive parameter needed to make the intensity values non-negative; some negative values might occur as an artifact of background correction. Because of the unknown shifts, the problem of estimating A , $f(t)$ and δ from the data is a non-linear PCA problem. To linearize the problem, we first perform a first-order Taylor expansion on (1) to give

$$S(t) = \beta_0 + Af(t) - A\delta f'(t). \quad (2)$$

This suggests a 2-component decomposition of $S(t)$. Thus, the first principal component (PC) of the spectral data (PC_1) provides a template for the peak shape. The second PC (PC_2)

captures the remaining signal-associated PC due to the shift and the remaining PCs are associated with noise [14]. Therefore, $S(t)$, can be modeled *via* the first two PCs.

$$S(t) = \beta_0 + \beta_1 PC_1(t) + \beta_2 PC_2(t). \quad (3)$$

By equating the terms in (3) with (2), we get $A = \beta_1$ and $\delta = -\beta_2/\beta_1$. Therefore, β_1 in (3) is constrained to be non-negative, while δ in (3) is constrained to avoid mixing peak shapes from neighboring regions. We estimate β_0 to be the smallest non-positive intensity in the spectrum and obtain estimates for β_1 and β_2 by treating (3) as a least-squares problem with constrained parameters. However, this approach sometimes does not align the spectra well, so we improve upon the final estimate with a direct least-squares fit in Step 2. Therefore, the role of Step 1 is to extract a template *via* PCA.

In summary, Step 1 consists of the following process:

- 1.1 Restrict the spectra to the m/z region of a peak.
- 1.2 Perform PCA on the restricted spectra.
- 1.3 Estimate A and δ in (1) *via* (3), and correct the misalignment in the restricted spectra.
- 1.4 Iterate Steps 1.2 and 1.3 until the mean squared difference or percent of variation explained by $PC1$ between templates of consecutive runs are within some tolerance level.

For pathological cases, when the templates do not contain a peak, the algorithm stops and returns the maximum intensity value and the corresponding location as the estimated peak height and location.

Step 2: Fitting of the template to the other spectra

From Step 1, we obtain the template $f(t)$ that we fit to the spectra individually. The fitting is done by minimizing the weighted mean squared error (MSE) between the template $f(t)$ and the measurements $S(t)$:

$$MSE = w \sum_t \{S(t) - \beta_0 - Af(t - \delta)\}^2/n, \quad (4)$$

where the weight w is defined as the reciprocal of the median intensity of the spectrum for the region and n is the number of points in the template. The MSE can also be used to compare the quality of the fit. For example, from the collection of all the spectra, we flag the spectral peaks with the top 2.5% MSEs as doubtful and set the parameter values to missing.

From the data it is obvious that some spectra have different peak shapes from the template. This might be expected because the peak shape are an aggregation of the template perturbed around the m/z of the protein: $S(t) = \beta_0 + \sum_{k=1}^d A_k f(t - \delta_k)$. Re-formulating this as a mixture model problem, we get:

$$\tilde{S}(t) = \sum_{k=1}^d \pi_k \tilde{f}(t - \delta_k), \quad (5)$$

where

- $\tilde{S}(t)$ is $S(t) - \beta_0$ divided by $\sum_t [S(t) - \beta_0]$ to be a probability mass function (pmf),
- $\tilde{f}(t)$ is $f(t)$ divided by $\sum_t f(t)$ to be a pmf,
- π_k 's are the mixing probabilities associated with A_k 's (i.e. $A_k = \pi_k \sum_t [S(t) - \beta_0] / \sum_t f(t)$) but restricted to be within 0 and 1, and
- d is the number of perturbation around the m/z of the protein.

In this context it is natural to define the amplitude of an aggregate as the sum of the individual amplitudes, i.e. $A = \sum_{k=1}^d A_k$, but

$$\begin{aligned} \sum_{k=1}^d A_k &= \sum_k \left\{ \pi_k \sum_t [S(t) - \beta_0] / \sum_t f(t) \right\} \\ &= \sum_t [S(t) - \beta_0] / \sum_t f(t), \end{aligned} \quad (6)$$

because $\sum_k \pi_k = 1$. So, no re-estimation is needed for the mixture model and we can use (6) to refine the estimate of the amplitude.

In summary, Step 2 consists of the following process:

- 2.1 Locate the valleys and peaks in the template with the modified SPF (see Supplementary material). For those templates with more than one peak, select the peak with the largest height and its valleys.
- 2.2 Restrict the value of δ to the range of whole numbers that confine the potential peak location of the spectrum to be fitted within the m/z range of the peak region.
- 2.3 For each eligible δ value, the A value can be computed *via* the mixture model approach (6). Compute the MSE.
- 2.4 The δ and A values with the smallest minimized MSE are $\hat{\delta}$ and \hat{A} .

3 Results

When we applied the ARS algorithm as outlined above to the clinical data set described in Section 2.1, we obtained 102 clusters from RS at a cutoff level of 5% FDR; during ARS, we found 151 peaks in these clusters. Among the 151 peak regions, only 7 of them had templates without a peak. Table 1 summarizes cluster sizes and number of peaks per cluster. Note that most of the unextended clusters are small (75 or 74% comprise three windows or less, corresponding to 15 measurements or less) and contain only one peak (73 or 72%, with a maximum number of six peaks for one cluster).

3.1 Performance of ARS templates in capturing the signal

Overall, the template approach implemented by ARS captures peak shapes well and provides good estimates of peak intensity and location. The results are best for strong signals

Table 1. Distribution of the size (in windows of 5 measurements each) and number of peaks detected for all 102 clusters found in the lung cancer data.

Windows <i>per cluster</i>	No of clusters
1 window	37 clusters
2 windows	24 clusters
3 windows	14 clusters
3 windows	6 clusters
5 windows	4 clusters
6 windows	7 clusters
27 windows	2 clusters
8 windows	2 clusters
9 windows	3 clusters
10 windows	1 cluster
11 windows	2 clusters
Peaks <i>per cluster</i>	No. of clusters
1 peak	73 clusters
2 peaks	16 clusters
3 peaks	9 clusters
4 peaks	2 clusters
5 peaks	1 cluster
6 peaks	1 cluster

and simple one-peak situations. We illustrate the performance of the template approach graphically and demonstrate that the results are satisfying even for clusters that are chosen to give ARS a hard time.

Based on the observed frequencies of peaks shown in Table 1, we decided to compare one-peak and two-peak clusters only. Visual inspection suggested furthermore that a maximum intensity of 20 or more for the median of the top third of the estimated peak intensities across spectra constituted a good cutoff for separating clusters with high and low signal strength. We therefore divided all one- and two-peak clusters into four groups: (i) one-peak clusters with strong signal, (ii) one-peak cluster with weak signal, (iii) two-peak clusters with strong signal, and (iv) two-peak clusters with weak signal. From each of the four groups we then selected as representative cluster the one whose average MSE (4) across all spectra was the 75th-percentile in the corresponding group. Using MSE as a measure for goodness of fit, where smaller values indicate better fit, this is a very conservative choice: we choose clusters whose fit to the ARS template is average in the worse half of the corresponding group. Since MSE was unavailable for peak regions that had templates without peaks, we adjusted the number of peaks down for those affected clusters in this subsection only.

Figures 1 (a)–(d) show the representative clusters of each group. Note that the plots show all 32 spectra simultaneously, but with special scaling so as to avoid visual clutter and to concentrate on the essential feature. For this purpose, we sorted the spectra for each cluster according to their estimated peak intensity and divided them into three groups of equal size, corresponding to spectra with high, medium, and low intensity (shown as dotted lines for low, dashed lines for

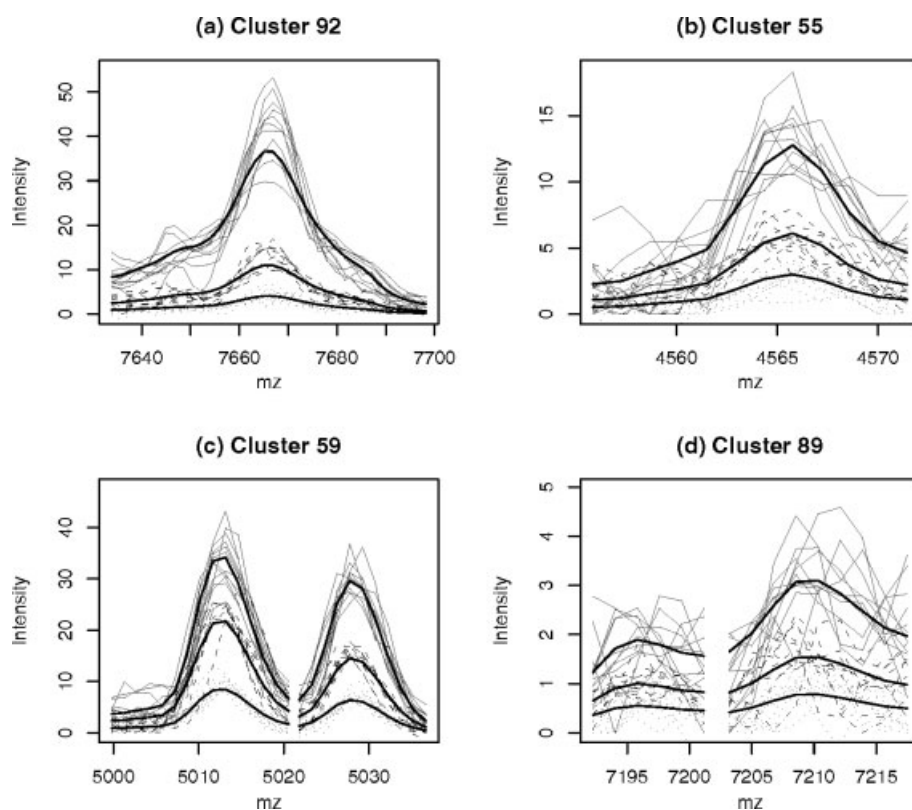


Figure 1. Illustration of the fit of the template peak shapes to clusters with (a) strong signal and single peak, (b) weak signal and single peak, (c) strong signal in one of two peaks, and (d) weak signal in two peaks. Spectra are categorized based on their estimated peak intensity as low intensity (dotted gray lines), medium intensity (dashed gray lines), and high intensity (solid gray lines); within each group, the spectra are scaled to the groups median estimated peak intensity, and the equally scaled template peak shapes are shown as overlaid solid black lines.

medium, and solid lines for high intensity in Fig. 1). In addition, the spectra in each group were all scaled to the group's median estimated peak intensity; the template spectrum for each cluster was also scaled to the median estimated peak intensity in each group, and is plotted as bold solid reference line for each group.

The fit is best for the strong signal situation in both (a) and (c), and it is worse in the weak signal for one peak situation in (b). The least satisfying fit is in the weak signal at two peaks situation in (d), where the largest scaled intensity for the spectra is five, indicating it is a low signal region. Further investigation suggests that those spectra with intensity greater than five had a good fit with the template while those below five are likely peaks due to noise. Note, however, that even in this noisy region, where the low and medium intensity spectra are basically pure noise and show no peak at all, the ARS template is flexible enough to provide estimates for the intensity and location; see Section 3.3 for a detailed illustration.

3.2 Comparison of ARS and standard method in peak detection

As Table 2 shows, ARS identified 151 peak regions, whereas the standard method only flagged 89 peak regions. We found that 78 peak regions identified by the standard method overlapped with 83 peaks identified by ARS. This disagreement is due to the fact that five (single) peak regions flagged by the standard method overlapped with two peaks from multiple peaks clusters suggested by ARS; the two peaks were typically located close to each other, and the standard method can get confused in the presence of multiple peaks, as illustrated in Section 3.3.

Table 2. The number of peak regions identified by ARS and the standard method in the lung cancer data. Overlapping peak regions are those identified by both methods, whereas non-overlapping regions are unique for each method.

	ARS	Standard Method
Overlaps	83	78
Non-overlaps	68	11
Total	151	89

Additionally ARS identified 68 peak regions that did not overlap with any peaks from the standard method, and conversely, 11 peaks were uniquely found by the standard method. We investigated all 79 mismatched peak regions visually and verified that (i) 60 out of 68 ARS peaks and (ii) all 11 standard peaks were plausible. We can use the McNemar test [15] to assess the agreement between the two methods in correctly identifying regions with or without peaks. This test only compares the discordant results, *i.e.* the 11 correct calls by the standard method that were mis-

sed by ARS, *vice versa*, the 60 correct calls made by ARS, plus its eight wrong calls. The McNemar statistic is therefore calculated as

$$\chi^2 = (11 + 8 - 60)^2 / (11 + 8 + 60) = 21.3,$$

which under the null distribution of equal power to assess peak regions follows a χ_1^2 distribution. The p-value for our data is highly significant at $p = 4.0 \times 10^{-6}$. This result demonstrates how ARS can, at the expense of a modest false discovery rate, identify significantly more verifiable peaks than the standard method. The attempt to reduce false positives using the standard method with the settings described in Section 2.2 has reduced the sensitivity of the standard method substantially. It is of course possible to increase the sensitivity, but our previous results [12] show that this will come at the price of a higher number of false positive peaks.

Note also that several of the 11 peak regions detected by the standard method and missed by ARS had very similar peak intensities across all spectra. It is an inherent feature of our signal detection algorithm to filter out peaks that have no significant variation between spectra. While these homogeneous peaks may represent bona fide proteins, they are found at almost constant concentration in all samples and are therefore unlikely to provide useful biomarkers. This should be seen as an advantage of ARS.

3.3 Comparison of ARS and standard method in peak quantification

We restricted our comparison of peak quantification to the 83 peak regions identified by ARS that overlapped with those found by the standard method, see Table 2. The comparison is based on the two crucial measurements provided by peak quantification: the estimated peak height or intensity, and the peak location or m/z value. Although both methods use different intensity profiles to obtain the estimated peak intensities, we can assess the strength of their agreement by calculating for each peak the Spearman rank correlation coefficient between ARS and standard intensities across all spectra. For the estimated peak m/z -values, we computed the difference between ARS and the standard method location for each spectrum, expressed as percentage of the smaller m/z -value; we used the median of these percentages to quantify the average shift in location between ARS and the standard method for a specific peak region.

Figure 2 shows the scatter plot of the correlations of the peak intensities *versus* the median percentage difference of the peak m/z -values, where every point represents one peak region. We have divided the scattered regions somewhat randomly into four groups labeled A–D in Fig. 2. About 84% of the peak regions have strong agreement in intensity; this corresponds to groups A and B, with correlations above 0.7. About 76% of the peak regions have strong agreement in location; this corresponds to groups A and C, with median

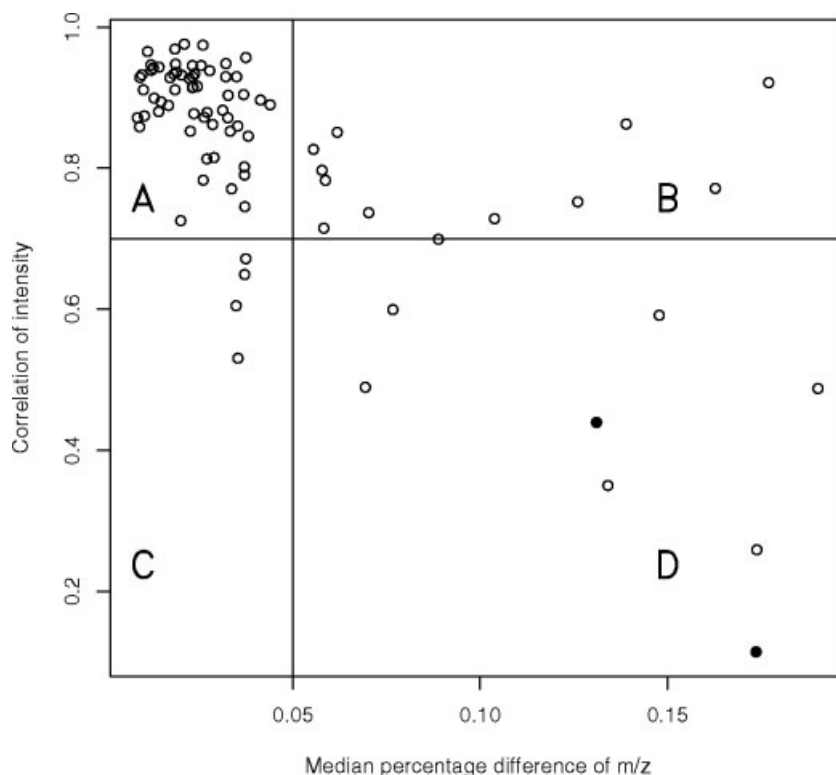


Figure 2. The scatter plot of the Spearman rank correlation of estimated peak intensities as estimated by ARS and the standard method versus the corresponding median percentage difference of the estimated peak m/z -locations for each peak region. Each point represents a peak region identified by ARS that overlaps with a peak region found by the standard method. The solidly marked point in the centre of D represents Cluster 39, for which Figure 3 shows more detail while the bottom right solidly marked point represents Cluster 45, for which Figure 4 shows more detail. We have divided the peak regions into four groups A–D, based on two criteria: correlation values ≥ 0.7 and median percentage difference values $\leq 0.05\%$.

percentage difference values below 0.05%. Altogether, 71% of the peak regions (group A) show strong agreement in both intensity and location.

We illustrate the performance of both methods in detail using representative peak regions from group D, *i.e.* where the methods disagree. We have chosen peak region at Cluster 39 (marked as solid symbol in Fig. 2 at the centre of region D), because its distance from the upper left corner in the scatter plot (representing perfect agreement between ARS and the standard method) is the median distance in group D, so they represent an average bad agreement. Both ARS and the standard method identified Cluster 39 as one peak. We have also chosen a peak region at Cluster 45 that corresponds to the point at the bottom right corner (solid symbol) to illustrate the reason why two ARS peaks overlapped with one standard method peak. For this cluster, ARS identified three peaks while the standard method identified two peak.

Figures 3 (a) and (c) for Cluster 39 show both ARS and the standard method identifying the same peak. The broken lines in the plots are the observed intensity spectra; the bold solid lines in the upper plots are the templates as fitted by ARS to the individual spectra, and the bold solid vertical lines with solid circles are the estimated peak intensities and m/z -values for both ARS (top) and the standard method (bottom). Surprisingly, in Fig. 3 (b), the standard method missed a clear peak that ARS identified and picked the slope as a peak instead. However, with fur-

ther examination of the cluster, we noted that the cluster could be more complex. From Fig. 3 (c) and especially Fig. 3 (d), the spectrum could be a two-peak cluster. In Fig. 3 (d), both methods selected a different peak. This further reinforces the difficulty of peak identification in SELDI as each spectrum may have different peak shapes in the same cluster and coupled with misalignment the complexity of peak identification increases. Despite of these issues, from this cluster, we see that ARS is potentially more robust and consistent in peak detection than the standard method.

Figures 4 (a) and (b) illustrate a typical situation where ARS performs better than the standard method; they also demonstrate how the discrepancy between 83 overlapping regions for ARS and 78 overlapping regions for the standard method discussed in Section 3.2 can arise. In Fig. 4 (a), both methods identified the left and right most peaks but ARS identified the shoulder between these two peaks that the standard method missed. However, in Fig. 4 (b), both methods identified the shoulder and right most peak but ARS identified the left most peak that the standard method missed. Although the standard method identified two peaks, in a situation where multiple peaks are in close vicinity of each other, the standard method gets confused and picks either one of them in different spectra. In comparison, ARS is more robust and manages to compensate for the misaligned m/z -values that cause the standard method to switch between neighboring peaks.

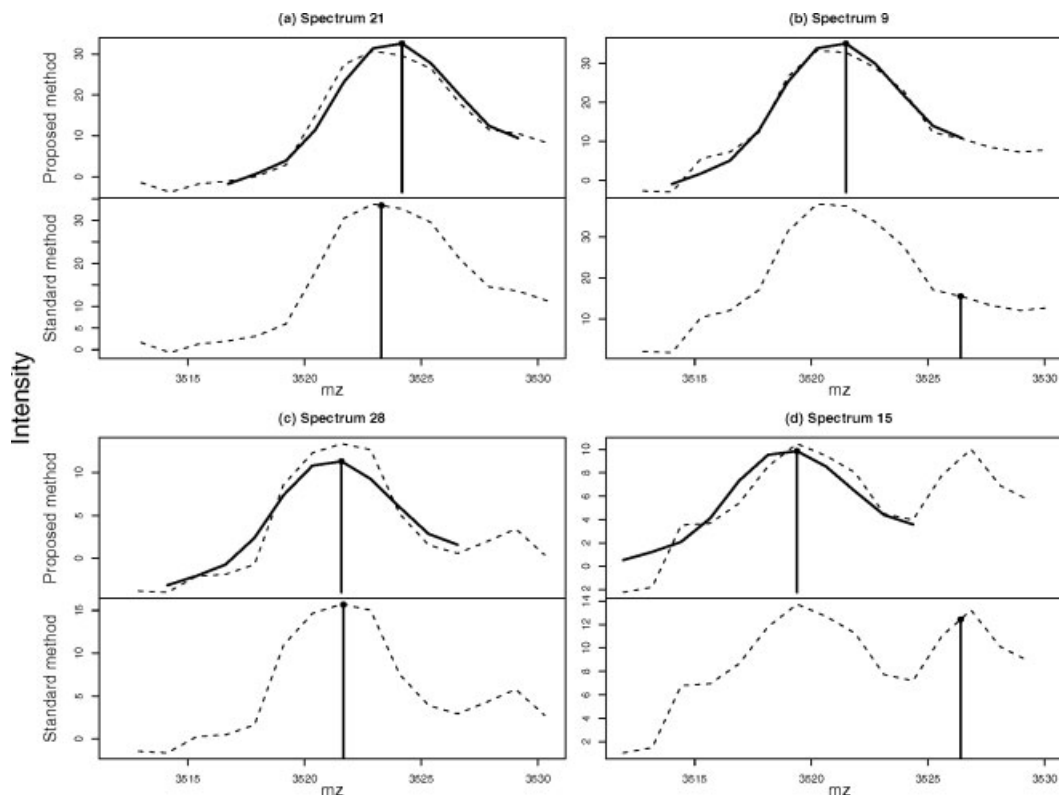


Figure 3. (a)–(d) Peak quantification of Cluster 39 in Fig. 2 for selected spectra. The top plot for each figure shows the observed intensities (broken lines) in the spectral region used by ARS, with the fitted peak shapes (bold solid lines) overlaid, and the vertical lines with solid circles indicating estimated peak intensities and m/z -values. The bottom plot shows the observed intensities (broken lines) in the spectral region used by the standard method, with the vertical lines with solid circles indicating its estimated peak intensity and m/z -values. Both methods identify one peak in this region. Both methods quantified the peak correctly in (a) and (c), but only ARS recognizes the peak in (b). The spectrum in (d) could contain two peaks with both methods identifying different peaks.

The poor performance by the standard method above might be expected given the difficult situation. However, even in a simple situation, it missed an obvious peak on the left of the spectrum in Fig. 4 (c) that ARS identified. From the plot, it is likely that the shoulder and the right most peak were not present in this spectrum and both methods returned low intensity values for them.

Finally, under severe misalignment situations, ARS may not perform well but still better than the standard method and this is illustrated in Fig. 4 (d). Both methods identified the left most peak in Fig. 4 (d) but both disagree on the peaks on the right. The standard method picked a trough, which is where the right most peak is located in the other three spectra. ARS identified the second peak, but the third peak is missed because of poor fit. On further investigation, we observed a peak after 3840 Da, suggesting the spectrum is misaligned (around 5 Da off) and this was corroborated with the duplicate of Spectrum 32. The duplicate was not severely misaligned and both methods identified the two peaks on the right, while only ARS identified the left most peak. This demonstrates the standard method being confused under severe misalignment resulting it to

identify a trough in (d), while ARS correctly identified the second peak but missed the third.

In summary, we have found that for the discordant peak regions in group D of Fig. 2, ARS has generally the advantage over the standard method, i.e. the situation in Fig. 4(a) is much more common than Fig. 4(c).

4 Discussion

For any proteomics approach, sensitivity and quantification are key issues in obtaining the high quality data required for the discovery of low-abundance peptides and proteins. We have developed ARS with the specific goal of providing sensitive signal detection and accurate peak annotation for weak but clearly biologically derived signals in spectral proteomics data. We have demonstrated that ARS can detect over 50% more peaks from lung cancer serum spectra than the standard CIPHERGEN method. Furthermore, we have shown improvements in peak annotation in ARS that can potentially benefit the downstream data analysis in biomarker research.

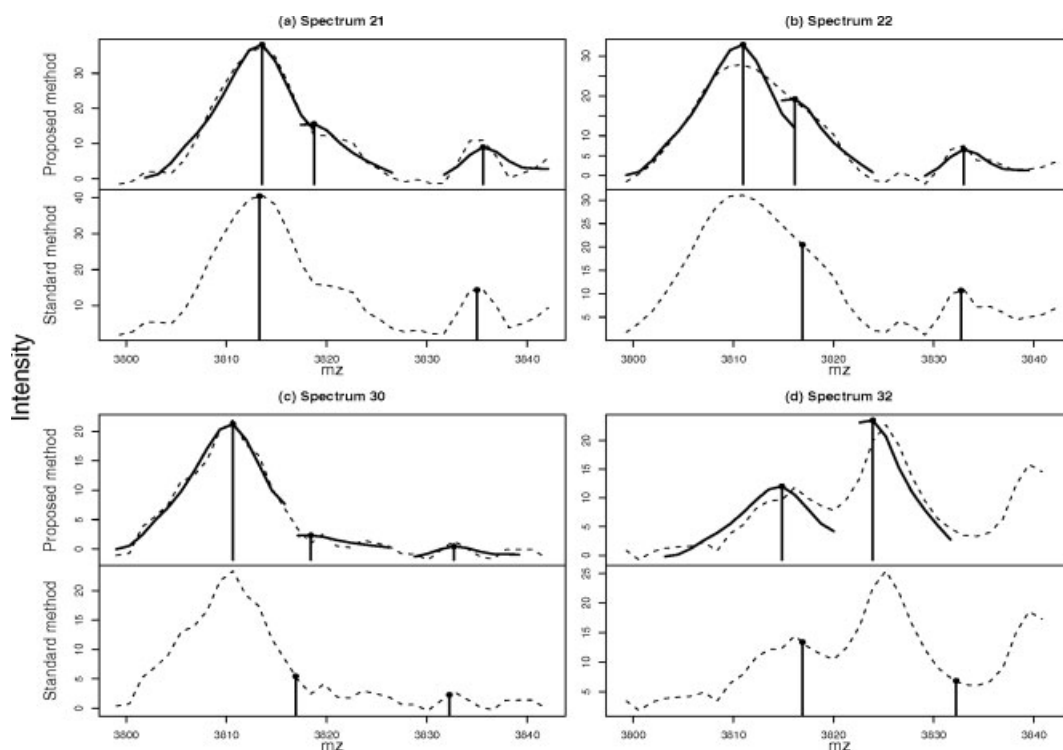


Figure 4. (a)–(d) Peak quantification of Cluster 45 in Fig. 2 for selected spectra. The top plot for each figure shows the observed intensities (broken lines) in the spectral region used by ARS, with the fitted peak shapes (bold solid lines) overlaid, and the vertical lines with solid circles indicating estimated peak intensities and m/z -values. The bottom plot shows the observed intensities (broken lines) in the spectral region used by the standard method, with the vertical lines with solid circles indicating its estimated peak intensity and m/z -values. ARS identifies three peaks in this region, while the standard method identifies only two peaks. (a) and (b) illustrate how the standard method can be confused by two neighboring peaks. (c) compares the performance of both methods in a simpler situation while (d) illustrates the performance of both methods under severe misalignment.

Linear TOF-MS data offers superior sensitivity compared to other methods, but inherent problems with mass accuracy make it difficult to label peaks accurately across multiple spectra. The template method implemented by ARS improves the interpretation of this type of raw spectral data; the scaling factor associated with the template fit allows the annotation of peaks that are partially merged, a frequent and challenging situation when measuring proteins in crude biological samples.

The improvements provided by ARS could potentially enable scientists to exploit data with low signal-to-noise ratio without requiring the same extensive manual validation as current methods. The performance of ARS's peak annotation, illustrated in our paper, could also reduce the amount of noise introduced into the subsequent search for single and multiple biomarkers in clinical data sets.

4.1 Advantages gained with ARS

The workflow of the Ciphergen standard method for detecting peak clusters can be outlined as follows:

1. Run a peak-finding algorithm at a conservative signal-to-noise ratio on all spectra separately.

2. If a minimum percentage of spectra have peaks from step 1 in a m/z region, generate a peak cluster in this region.
3. For all spectra that do not have a peak in the current region from step 1, run the peak-finding algorithm again, with the signal-to-noise requirement relaxed.
4. For all spectra that do not have a peak in the current region from step 1 or step 3, select a small peak in the region arbitrarily.

ARS has a clear advantage in the peak detection (step 1 and 2), because it uses all spectra simultaneously to identify peak regions. The standard method is also at a disadvantage in regard to peak annotation because of steps 3 and 4, which try to quantify peaks in low-intensity spectra without making use of cross-spectral information.

4.2 Comparison with other peak detection procedures

ARS combines threshold-based and template-based peak detection. Simple threshold-based peak-finding algorithms try to identify noise levels so that all measurements above a

specific noise-level qualify as peaks, *e.g.* [8, 9]. This approach works well for clear peaks with strong signal, but it is inadequate for peaks with low intensity or unclear definition. Due to calibration and scaling issues it would also be difficult to define a noise level that could be applied across different spectra, so threshold-based methods are generally single-spectrum. We have, instead, identified *regions found to contain significant variation between spectra*; in this way, we can guarantee that (i) there is a reasonable chance that the region actually contains potential peaks for biomarker research (depending on the cutoff criterion during signal detection), and (ii) we apply the SPF to the spectrum with the strongest peak/signal, *i.e.* we assure that the SPF is used under the most favorable conditions.

Template-based methods on the other hand try to detect the same specified peak shape at a given *m/z* location in all spectra. This has the advantage that some specified structure can be imposed across noisy spectra, but it begs the question of how to define the peak template. Traditional approaches assume some kind of parametric shape for peaks, *e.g.* [16]. Recently, there have been attempts to estimate the template more directly from the data, *e.g.* [17] in the context of generic chromatography, where the authors propose using an application-dependent standard peak, derived as the average of multiple peaks of multiple hand-annotated spectra. Based on our experiences with the variability of protein peaks, however, we are not convinced that such an approach would be flexible enough to work with SELDI data. Our approach applies PCA across all spectra and takes into account misalignment in the spectra when extracting a template for each peak region. So the template from this approach is automatically the best fitting shape for the collection of spectra in the region.

An interesting multi-spectral method called SSA (simultaneous spectrum analysis) that shares some features with ARS has been described recently in [10]. The authors also propose a true multi-spectral approach based on (unmodified) F-statistics that outperforms the standard method in terms of peak detection and quantification. The crucial difference between SSA and our approach lies in the fact that the F-statistics in SSA are based on the biological grouping of the spectra, *e.g.* knock-out *vs.* normal mice or benign *vs.* cancerous prostate tissue. Although the ultimate interest is in the between-class variation, very few markers will be significantly different between classes, so if we test every region across the range of spectra without any screening, the potential for false discoveries is large and it is harder to detect the real signal.

4.3 Summary

We have combined a peak quantification algorithm with a signal detection algorithm to address the low specificity and poor peak quantification of the standard peak detection method. We have demonstrated that our proposed method can outperform the standard method in several aspects: (i) ARS captured several peak regions in the spectral data that were missed by the standard method, (ii) ARS was better at classifying regions as peak or non-peak regions, (iii) ARS was less easily confused by two or more closely neighboring peaks and/or *m/z*-misalignment than the standard method, and (iv) the ARS results appeared to quantify the peaks better than the standard method.

5 References

- [1] Zhang, Z., Bast Jr., R., Yu, Y., Li, J. *et al.*, *Cancer Res.* 2004, *64*, 5882–5890.
- [2] Wadsworth, J., Somers, K., Cazares, L., Malik, G. *et al.*, *Clin. Cancer Res.* 2004, *10*, 1625–1632.
- [3] Kozak, K., Amneus, M., Pusey, S., Su, F. *et al.*, *Proc. Natl. Acad. Sci. USA* 2003, *100*, 12343–12348.
- [4] Hutchens, T., Yip, T., *Rapid Commun. in Mass Spectrom.* 1993, *7*, 576–580.
- [5] Fung, E., Enderwick, C., *Computational Proteomics Suppl.* 2002, *32*, S34–S41.
- [6] Malyarenko, D. I., Cooke, W. E., Adam, B.-L., Malik, G. *et al.*, *Clin. Chem.* 2005, *51*, 65–74.
- [7] Coombes, K., Tsavachidis, S., Morris, J., Baggerly, K. *et al.*, *Proteomics* 2005, *5*, 4107–4117.
- [8] Yasui, Y., Pepe, M., Thompson, M., Adam, B. *et al.*, *Biostatistics* 2003, *4*, 449–463.
- [9] Coombes, K., Fritsche Jr., H., Clarke, C., Chen, J.-N. *et al.*, *Clin. Chem.* 2003, *4*, 1615–1623.
- [10] Carlson, S., Najmi, A., Whitin, J., Cohen, H., *Proteomics* 2005, *5*, 2778–2788.
- [11] Birkner, M., Hubbard, A., van der Laan, M., Skibola, C. *et al.*, *Stat. Appl. Genet. Mol. Biol.* 2006, *1*, Article 11.
- [12] Tan, C., Ploner, A., Quandt, A., Lehtiö, J. *et al.*, *Bioinformatics* 2005, *22*, 1515–1523.
- [13] Anderson, T., *An introduction to multivariate statistical analysis*, 2nd edition, Wiley, New York, 1984.
- [14] Stoyanova, R., Kuesel, A., Brown, T., *J. Magn. Reson., Ser. A* 1995, *115*, 265–269.
- [15] Lachenbruch, P., *Encyclopedia of Biostatistics: McNemar Test*, Wiley, New York, 1998.
- [16] Danielsson, R., Bylund, D., Markides, K., *Anal. Chim. Acta* 2002, *452*, 167–184.
- [17] Steffen, B., Müller, K., Komenda, M., Koppmann, R. *et al.*, *J. Chromatogr. A* 2005, *1071*, 239–246.