analytical chemistry

ARTICLE

pubs.acs.org/ac

Pinpointing Biomarkers in Proteomic LC/MS Data by Moving-Window Discriminant Analysis

Tom G. Bloemberg,[†] Hans J. C. T. Wessels,[‡,§] Maurice van Dael,[‡] Jolein Gloerich,[‡] Lambert P. van den Heuvel,[‡,∥] Lutgarde M. C. Buydens,[*,†] and Ron Wehrens[*,†,^]

[†]Radboud University Nijmegen, Institute for Molecules and Materials, Heyendaalseweg 135, 6525 AJ, Nijmegen, The Netherlands

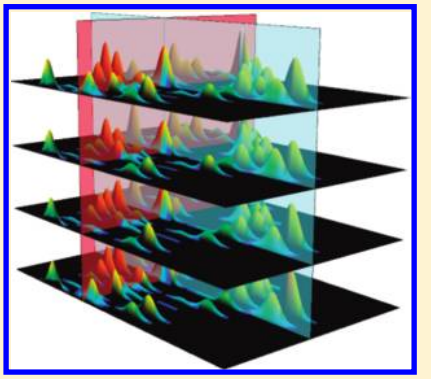
[‡]Nijmegen Proteomics Facility, Department of Laboratory Medicine, Laboratory of Genetic, Endocrine and Metabolic Diseases, Radboud University Nijmegen Medical Centre, The Netherlands

[§]Nijmegen Centre for Mitochondrial Disorders, Department of Laboratory Medicine, Laboratory of Genetic, Endocrine and Metabolic Diseases, Radboud University Nijmegen Medical Centre, The Netherlands

[∥]Department of Pediatrics, Radboud University Nijmegen Medical Centre, The Netherlands

Supporting Information

ABSTRACT: The identification of differential patterns in data originating from combined measurement techniques such as LC/MS is pivotal to proteomics. Although "shotgun proteomics" has been employed successfully to this end, this method also has severe drawbacks, because of its dependence on largely untargeted MS/MS sequencing and databases for statistical analyses. Alternatively, several MS-signal-based (MS/MS-independent) methods have been published that are mainly based on (univariate) Student's *t*-tests. Here, we present a more robust multivariate alternative employing linear discriminant analysis. Like the *t*-test-based methods, it is applied directly to LC/MS data, instead of using MS/MS measurements. We demonstrate the method on a number of simulated data sets, as well as on a spike-in LC/MS data set, and show its superior performance over *t*-tests.



Liquid chromatography/mass spectrometry (LC/MS) has become the de facto standard as a platform for proteomic biomarker discovery studies.[1,2] In a typical proteomic biomarker search, samples originate from a tissue biopt or bodyfluid (plasma, urine, cerebrospinal fluid) of two groups: a case group and a control group.[3,4] Currently, the most important methodology for finding systematic differences between these groups of samples by means of LC/MS is shotgun proteomics.[5,6] This methodology comprises MS/MS sequencing measurements in combination with database searching to identify proteins in the sample, after which the abundances (based on MS signal intensities or spectral counts) of the identified proteins are compared. There are intrinsic problems associated with this approach: the MS/MS sequencing is largely random, and quantitation of peptides is dependent on their identification.

The sequencing problem is immediately obvious from the term "shotgun". Although peptide sequencing by MS/MS is often referred to as data-dependent, this means, in practice, that only the most intense peptide signals in a certain retention time range will be selected for sequencing. Thus, many peptides will not be sequenced at all, and the choice of peptides to be sequenced is random with respect to the problem under investigation. Moreover, when multiple samples are considered—as in a biomarker search—there is no assurance that a protein that is present in certain samples will actually be detected in all of them.

On top of the limited number of sequencing events, a general problem that any MS/MS-based biomarker search strategy faces is its dependence on peptide identifications and quantitations.[6] Any peptide ion not readily identified by database searches (typically about 70−80% of the acquired MS/MS spectra after validation, in our experience) is therefore not quantified (and is thus neglected), even when it might be the most significant dissimilarity between sample sets. Alternative strategies for the identification of peptide sequences from fragment ion spectra, such as computer-assisted de novo sequencing or manual interpretation, might help to alleviate this problem. However, these methods also have a limited success rate and are extremely time- and computer-intensive. With the hardware and software currently available, it is therefore simply impossible to successfully identify the majority of peptides in a sample, thus presenting an intrinsic problem for MS/MS-driven quantitation strategies.

Received: February 8, 2011
Accepted: May 10, 2011
Published: May 10, 2011

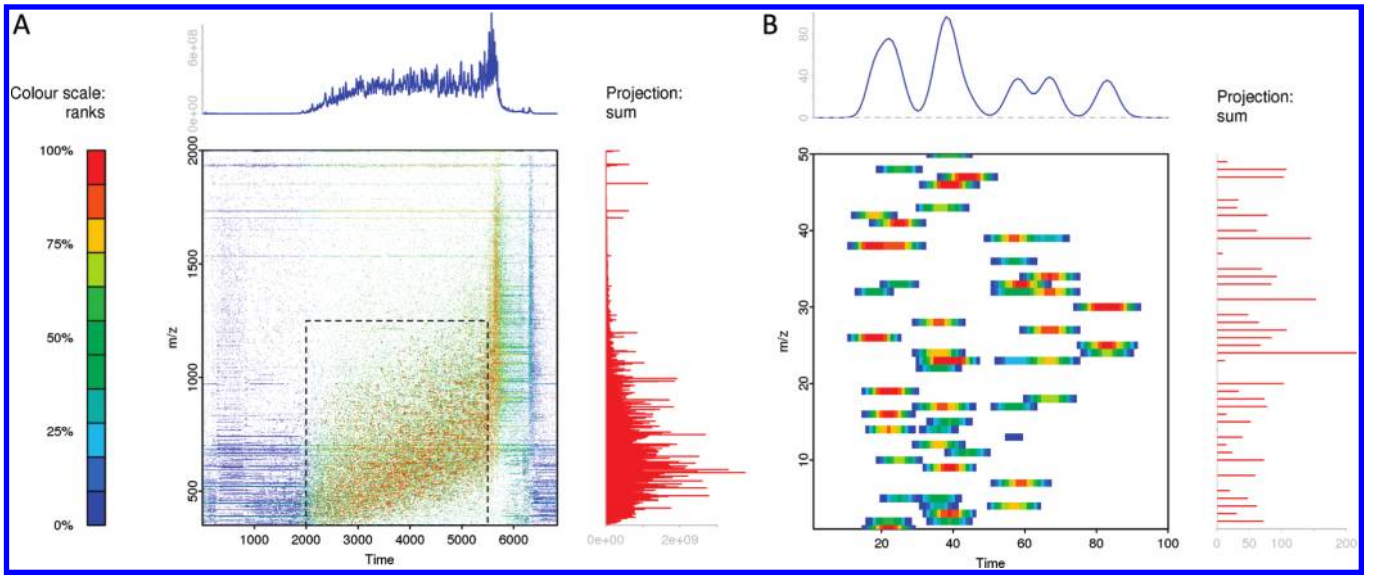


Figure 1. Images of the matrix representations of (A) the first LC/MS measurement of the *E. coli* data set and (B) the first sample from the illustration data set. Columns are mass spectra; rows are chromatograms. In both panels, the blue chromatogram at the top is the sum of all ion intensities at each given time point: the total ion current (TIC). The red mass spectra at the right are the total mass spectra. The color scales for the two panels are similar, but relative to the samples' intensity distributions. The rectangle in panel A indicates the part of the *E. coli* samples that was analyzed in this work.

For the above-mentioned reasons, there is an incentive for database-free and MS/MS-free analysis methods that are able to extract differential patterns from LC/MS data without the need for a peptide or protein identification step in the search process.[7] Several such methods have been described in the literature.[8−10] Commonly, they employ some modified version of a *t*-test to locate differential patterns in LC/MS data. For perfect data, with a large number of samples, no noise or intensity changes, and no misalignments in the chromatographical direction, these methods perform adequately. Unfortunately, real-life data are far from perfect: noise is omnipresent, and the requirement for many samples opposes that of no misalignments due to, for example, column replacements and variable performance of the ionization source, liquid chromatograph, and mass spectrometer over time.

In this article, we introduce the concept of moving-window discriminant analysis (MWDA). This approach uses a combination of principal component analysis (PCA[11−13]) and linear discriminant analysis (LDA[12−15]), collectively known as PCA−LDA[16−20] to compare complete mass chromatograms or mass spectra extracted from the LC/MS data rather than single intensities, as in *t*-tests. We show that, for perfect data, its performance is similar to that of multiple *t*-tests, but it is far more robust to imperfections, thus outperforming *t*-tests in realistic situations. We also apply both methods to a set of spike-in LC/MS data (the *E. coli* data set) that will be made publicly available as a benchmark model for complex samples with biomarkers.[21] We show that MWDA has a much steeper initial true positive rate than *t*-tests for these real data as well and that it is capable of identifying a significant number of spiked-in peptides in the data. We expect the method to be of use in both current bottom-up (peptide-based) and upcoming top-down (whole-protein-based) proteomics.[22,23]

THEORY

After the typical data processing steps described in the Supporting Information, an LC/MS sample can be represented as a matrix. Figure 1 shows matrix representations of sample 1 from the *E. coli* data set and sample 1 from a small noiseless data set that was simulated purely for illustrative purposes (the illustration data set). The columns of both matrices are mass spectra separated by (elution) time, and their rows are chromatograms distinguished by mass-to-charge (m/z) ratio. Generally, complete proteomic LC/MS studies consist of several tens of samples. Stacked on top of each other, their matrix representations form a three-dimensional array or data cube of dimensions (number of m/z values) × (number of time points) × (number of samples). Figure 2A shows such a data cube of samples 1−4 of the illustration data set; the horizontal slices are matrices like the one in Figure 1 representing the individual LC/MS measurements.

Multiple *t*-Tests. A basic univariate approach for finding potential biomarkers applies two-sample Student's *t*-tests[24] to the intensities of the case and control samples in the vertical columns (Figure 2C) at all time−m/z combinations in the data cube. Formally, when a set of multiple inferences is considered simultaneously, a multiple testing correction is indicated to prevent the null hypotheses from being rejected incorrectly too often. In the context of biomarker searches, however, it is not so much the particular p values for the individual hypothesis tests that are of interest, but rather their order:[10] the smallest p value is most likely to be caused by a true biomarker. Because multiple testing corrections are one-to-one functions of the input p values, their order does not change, and the correction can safely be left out.

Moving-Window Discriminant Analysis. Instead of just columns, it is also possible to take complete vertical slices from the data cube (Figure 2A,B). There are two options: slicing along the retention time axis gives matrices containing chromatograms (LC matrices, blue slice in Figure 2B), whereas slicing along the m/z axis results in a matrix of mass spectra (MS matrices, red slice in Figure 2B). These matrices are amenable to multivariate analyses, which are commonly employed for regular one-dimensional spectroscopical or spectrometrical data (e.g., infrared

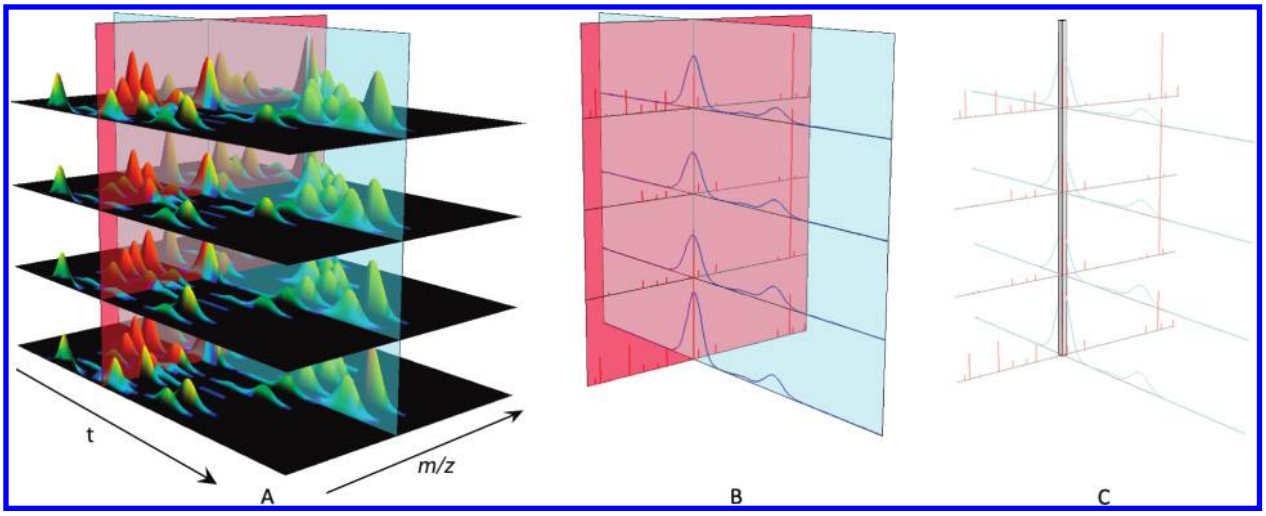


Figure 2. Part of the data cube of the illustration data set. The black horizontal slices in panel A represent the first four LC/MS samples of the data set. The blue vertical slice is an LC matrix, containing chromatograms; the red slice contains mass spectra and thus is an MS matrix. The vertical column in panel C contains a single intensity per sample. These intensities can be subjected to a *t*-test, whereas the LC and MS matrices are amenable to multivariate analyses such as LDA.

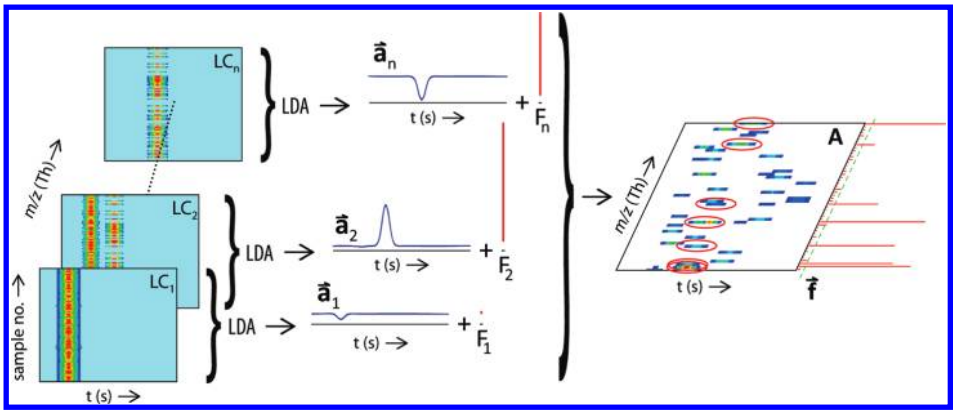


Figure 3. The MWDA procedure applied to LC matrices of the illustration data set: PCA−LDA is applied to the subsequent LC matrices of the small simulated data set. Each application results in a discriminant coordinate **a** (a vector) and a Fisher quotient *F* (a scalar). Combined, the Fisher quotients make up a vector **f** that can be plotted as a mass-spectrum-like structure, with the intensities representing the classifiability of the corresponding LC matrices. The discrimininant coordinates together make up a matrix **A** that is similar to the matrix of an LC/MS measurement (Figure 1). A high Fisher quotient means that the corresponding LC matrix is well-classifiable and, thus, is likely to contain one or more chromatographic profiles that distinguish between the sample classes. The profile of interest can be found from the associated discriminant coordinate.

spectra, NMR spectra, mass spectra, etc.) and chromatographic data.[16−20,25]

MWDA proper comprises the application of PCA−LDA (see the Supporting Information) to the subsequent mean-centered LC matrices in the data cube. Moving the window along the *m*/*z* axis or along the time axis, MWDA produces a scalar Fisher quotient *F* and a discriminant coordinate vector **a** for each subsequent matrix (Figure 3). Combined, the Fisher quotients make up a vector resembling a mass spectrum. Complementary to the Fisher quotients, the individual discriminant coordinates resemble chromatograms. Taken together, the discriminant coordinates form a matrix **A** of dimensions (number of *m*/*z* values) × (number of time points), that is, resembling a sample from the original data set.

Potential biomarkers are now expected to be found in **A** at the *m*/*z* values (rows) specified by high values of the Fisher quotient and in the orthogonal time direction (columns) by the highest peak (absolute value) in the previously specified row. In the original samples, disregarding misalignments, the peak of the biomarker-peptide is located at that same position.

Here, we use the LDA Fisher quotients *F* to determine whether an LC matrix contains a biomarker signal. The percentage (of samples) correctly classified (PCC) in a cross-validation setting can be used to the same end. As opposed to *F*, which is on a continuous scale, the discrete nature of the computationally intensive PCC can obscure interesting differences between the LC matrices for small sample numbers, however. Apart from optimizing the settings of PCA−LDA in MWDA, it might also be worth considering other discriminant methods [e.g., partial least-squares discriminant analysis (PLS-DA), elastic nets[26]] in similar fashion. Regression methods [e.g., PLS, principal component regression (PCR)] would also be of interest to use in an analogous approach for data from time series or other settings that differ from the typical two-class case−control setup.

EXPERIMENTAL SECTION

Both simulated and real LC/MS data sets were used in this work. A short description of both types of data sets follows; more elaborate information is presented in the Supporting Information, including the functions and scripts that were used for simulating data.

Thirty-two data sets were simulated in R,[27] to assess the relative performances of MWDA and *t*-tests for different amounts of misalignment. All of these data sets had the same dimensions ($1000 \times 2000 \times 17$). The number of samples (17) and their class distribution (11 from class one, 6 from class two) were equal to those in the real *E. coli* data set. All class one samples in all 32 data sets contained 1000 compounds, characterized by a single Gaussian elution profile, but different numbers (between 1 and 10) of MS peaks. That is, the numbers of peaks may differ slightly between samples in data sets A and B but are the same for all samples in data set A. Compared to class one samples, class two samples contain five extra compounds that represent the biomarkers. The peak width along the m/z axis is a single data point; along the retention time axis, it gradually gets wider according to standard deviation (sd) = $\max[1, 1 + \log_{10}(t)]$; that is, the standard deviation for the Gaussian line shape goes from 1 at $t = 0$ to 4 at $t = 1000$.

The parameter of interest in the simulated data is the misalignment between identical compounds in different samples. To investigate its influence, the 32 simulations consisted of four groups (1−4) of eight data sets (A−H) each. Each group was essentially a copy of the other ones (i.e., the compounds in data set A of group 2 were exactly the same as those in data set A of group 1), but the groups differed in one main aspect: In group 1, identical compounds in all samples were located at the exact same positions along the elution time axis, whereas in groups 2−4, they had a random normal shift with standard deviations of 10, 50, and 100 data points, respectively.

The *E. coli* set is a real $1801 \times 2000 \times 17$ [(number of m/z values) × (number of time points) × (number of samples)] proteomics data set that consists of LC/MS measurements of tryptic digests[28] of *Escherichia coli* protein homogenate. Six of the 17 samples were spiked with bovine carbonic anhydrase. As opposed to the simulated data sets, there was no ground truth (i.e., prior knowledge about carbonic anhydrase peptides) available for this set; MWDA and *t*-tests results were assessed afterward, based on two further types of experiments: LC/MS/MS measurements and direct-infusion nano electrospray ionization (here called nanospray ionization, NSI) MS measurements of tryptically digested carbonic anhydrase only.

Data Analysis. *Simulated Sets: MWDA and t-Tests.* In principle, MWDA can be applied to either the LC matrices or the MS matrices of LC/MS data. Misalignments along the retention-time (LC) axis inhibit its application to MS matrices, however. Thus, MWDA was applied to only the LC matrices of each set. The performance of PCA−LDA is sensitive to the number of principal components retained prior to LDA. We therefore opted to use the first three data sets (A−C) out of each group of eight as a training set to determine a tentative optimum for the number of principal components to retain. The number of PCs was then fixed for the subsequent MWDA analyses of the five sets (D−G) with similar characteristics per group (the test sets). For each set, the 1000 Fisher quotients of the respective LC matrices were used to determine which matrices to designate as positives.

For comparison, two-sample, two-sided Student's *t*-tests were applied to the intensities at the 2×10^6 individual m/z−time points of each set, using pooled variances. The resulting 2×10^6 p values cannot be directly compared to the 1000 F values obtained by applying MWDA to the LC matrices. Therefore, to enable comparison with MWDA, the lowest p value out of the 2000 for each LC matrix was used to determine if the matrix was a positive according to the *t*-tests.

E. coli Benchmark Set: MWDA and t-Tests. After the data processing steps described in the Supporting Information, the LC matrices of the real spike-in set were analyzed using MWDA in two steps: First, the complete $1801 \times 2000 \times 17$ array was analyzed similarly to the simulated sets. The number of PCs to use could not be based on a training set with a known ground truth in this case and was therefore determined as the rounded average of the first "knees" or "elbows" in the scree plots[11] of 10 randomly chosen LC matrices, resulting in five PCs being retained for every LC matrix. Assuming equal null distributions for MWDA on all LC matrices, a permutation test of 18010 total permutations was performed by making 10 class permutations per LC matrix. The significance level was set at 0.10; that is, only Fisher quotients that were higher than the 90% lowest Fisher quotients of the permuted data sets were designated positive. Fifty of 1801 Fisher quotients were larger than the corresponding threshold. Put differently: 50 of 1801 null hypotheses of "no difference" were rejected. About half of the peaks were identified as corresponding to $(m + 1)/z$, $(m + 2)/z$, and $(m + 3)/z$ peaks of isotope patterns of which the monoisotopic peak was also identified. These isotope peaks were retained in the further analysis.

To increase the precision of the 50 m/z values, binning, as described in the Supporting Info, was performed again, starting all over from the LC/MS data in mzXML format and making 51 new bins of width 0.01 Th spanning the range of each original 0.5-Th-wide bin that was positive in the first analysis. The resulting 2550 new chromatographic matrices were subjected to MWDA again, and for each group of 51 sub-bins, the m/z value of the one with the highest Fisher quotient was noted. These 50 precise m/z values were subsequently used for comparison with the additional LC/MS/MS and direct NSI MS measurements (i.e., without chromatographic separation) of carbonic anhydrase only.

For the *t*-tests, a similar procedure was followed: Again, the complete *E. coli* array was analyzed similarly to the simulated sets. Rather than making use of the theoretical *t*-distribution, permutation tests were used here as well: the class labels were permuted 10 times, and after each time, the complete *E. coli* array was analyzed, using the permuted class labels. For both the permuted and original class labels, the lowest p value per LC matrix was determined, as for the simulated sets. Next, the positive LC matrices, with p values below the 0.10 significance level were determined, resulting in 85 positives. For these, the LC/MS data were rebinned and reanalyzed, and the precise results were again compared with those from the additional measurements.

Finally, after comparing the performance of MWDA and *t*-tests for the *E. coli* set, we decided to compare the performance of the *t*-tests with that of random m/z values, to assess the baseline level for the procedure. To this end, the entire *t*-test procedure was applied to the *E. coli* set with permuted class labels. Because a permutation test for permuted data would be meaningless, an arbitrary number of 72 p values were designated positive, corresponding to an integer percentage of 4% of the total number of p values and between the numbers of positives

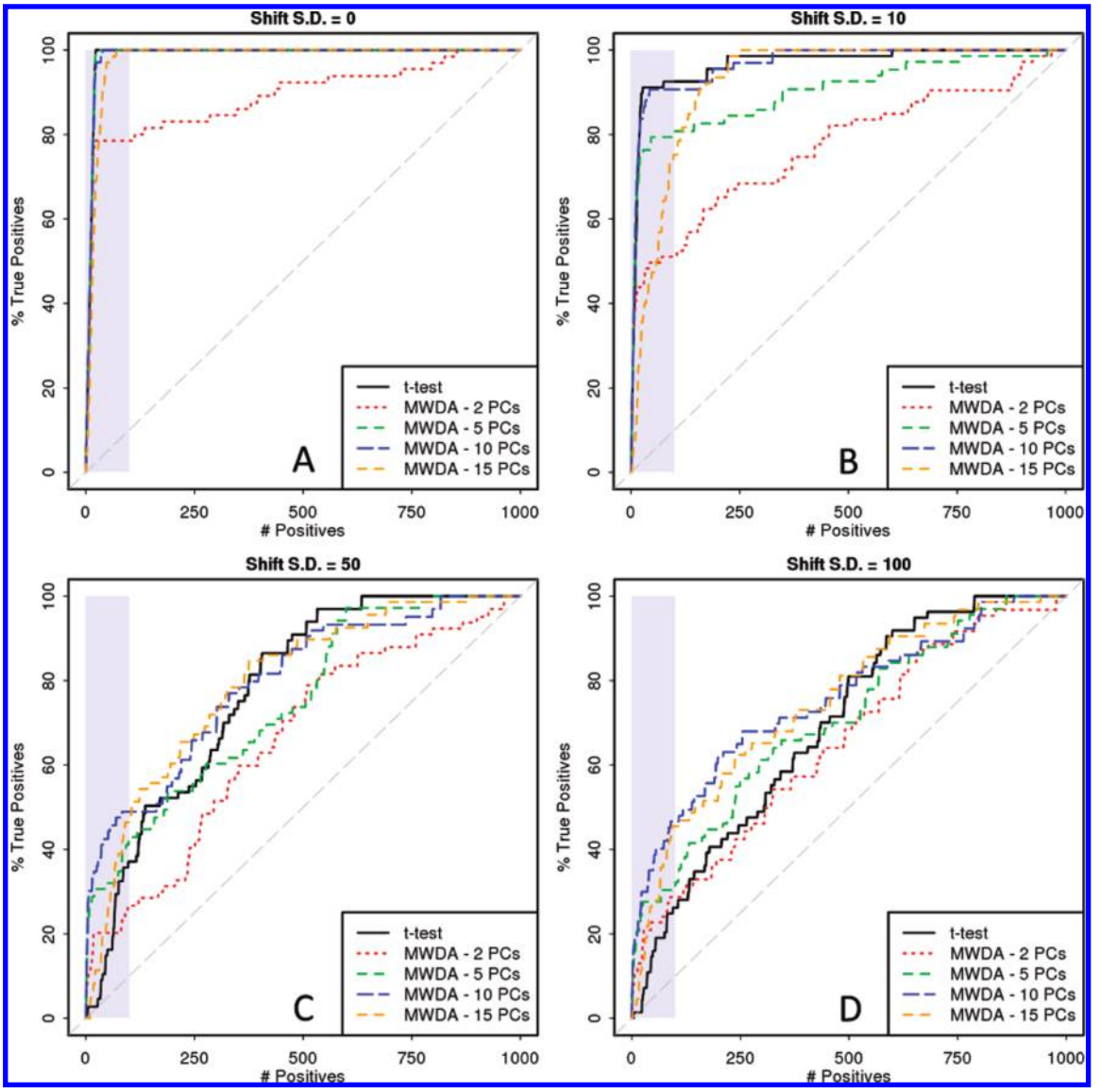


Figure 4. Average percentage true positive rates for the simulated training sets, using *t*-tests and MWDA. The shaded blue areas represent the interesting regions of the plots where up to 10% of the LC matrices were designated positive.

for MWDA and *t*-tests. The rest of the procedure was as described above for the *t*-tests on unpermuted data.

RESULTS

For a comparison of classifiers, it is good practice to draw receiver operating characteristic (ROC) curves. In short, an ROC curve is a plot of the true positive rate (the fraction of positives that are true positives) versus the false positive rate (the fraction of negatives that are false negatives). For a comprehensive explanation, the reader is referred to the excellent article by Fawcett.[29] A prerequisite for making an ROC curve is thus the identification of both positives and negatives as being true and false positives and negatives, respectively. Identification of the positives for the *E. coli* set is already a cumbersome task, but it is manageable in this particular case because the spiked protein is known and can be measured separately, after which its NSI MS or LC/MS signals can be compared with those on the limited list of positives. Identifying and quantifying the number of negatives in an objective manner is a problem of a different order, however, because it involves many more signals, often of lower quality, and the number of true and false negatives is very dependent on the exact criteria that are chosen. Therefore, for ease and objectivity, instead of ROC curves, we chose to plot the true positive rates versus the total number of LC matrices designated positive for the *E. coli* set, as well as, consequentially, for the simulated sets. For the latter, the two types of plots are nearly identical, as can be seen from comparing the ROC curves in the Supporting Information with the plots in the following sections, and in all cases discussed in this article, the plots convey the same message, namely, that the curve that shows the steepest ascent near the origin represents the best biomarker identification method.

Simulated Sets. Figure 4 shows plots of the average percentage of true positives versus the number of LC matrices designated positive for the training set. The curves were constructed in a fashion similar to the ROC curves,[29] by simply ordering the *F* or *p* values from high to low significance and comparing the associated LC matrices with the list of LC matrices known to

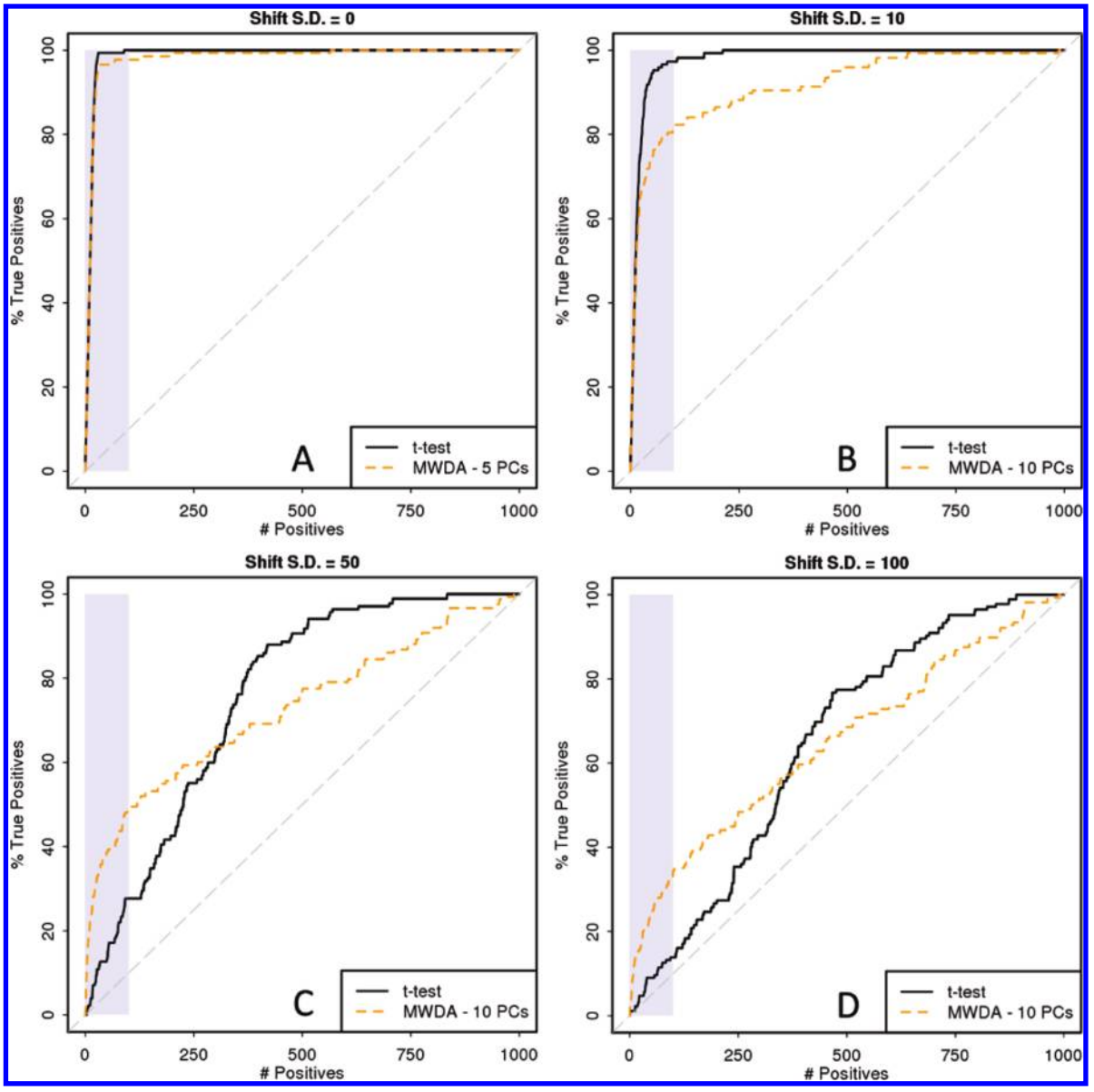


Figure 5. Average true positive rates for the simulated test sets, using *t*-tests and MWDA. The numbers of principal components used for MWDA are based on the results of the training sets. The shaded blue areas represent the interesting regions of the plots where up to 10% of the LC matrices were designated positive.

contain an isotope peak of an added component. Percentages rather than numbers were used, because the total number of true positives in each of the eight data sets varied, as components can have different numbers of isotope peaks. Averages were taken over the three data sets from each group with similar shifts. Based on Figure 4, the numbers of principal components retained for MWDA analyses of the test sets were 5 for the perfectly aligned set and 10 for the consecutive misaligned sets. The shaded blue areas are the arbitrary "regions of interest" where up to 10% of the LC matrices were designated positive. For ROC curves, the partial area under the curve (pAUC) could be calculated in the similar region.

Figure 5 shows the results of MWDA analyses of the test sets, using the numbers of PCs indicated by the training sets. It is clear that, for the perfect case of no misalignment, *t*-tests and MWDA perform likewise. For a small misalignment on the order of the peak width, the performance of the *t*-tests is hardly affected, and the performance of MWDA drops. For stronger misalignments comparable to the ones observed in the *E. coli* set, however, the performance of the *t*-tests drops rapidly and becomes significantly lower than that of MWDA.

***E. coli* Benchmark Set.** MWDA and *t*-tests were applied to the *E. coli* set blind to any ground truth. Subsequently, LC/MS/MS and direct NSI MS measurements of tryptically digested pure carbonic anhydrase were performed, and the results were compared with the MWDA and *t*-tests lists of positives. The final results of these comparisons are summarized in Table 1.

Two things are immediately clear from this table: (1) Both the real *t*-test and MWDA performed substantially better than the baseline level, given by the application to *t*-tests in the permuted-class situation, and (2) MWDA led to much better results than the *t*-tests; that is, even despite the much larger number

Table 1. Results for MWDA and *t*-Tests Applied to the *E. coli* Set

method	P^b	TP[c]	FP[d]	no. of CA ions[e]	m/z (NSI)[f]
MWDA	50	44 (88%)	6 (12%)	44 (88%)	41 (82%)
t-tests	85	28 (33%)	57 (67%)	25 (29%)	23 (27%)
t-tests (baseline[a])	72	9 (13%)	63 (88%)	8 (11%)	3 (4%)

[a] Data with permuted class labels. [b] Number of positives as determined by each method. [c] Number (percentage) of true positives. [d] Number (percentage) of false positives. [e] Total number of carbonic anhydrase peptide ions identified by LC/MS/MS. [f] Number (percentage) of m/z values for which a peak is identified in the direct NSI mass spectrum.

of positives for the *t*-tests, the number of true positives for MWDA was considerably higher. This fact is also clearly visible in Figure 6, which shows the initial true positive rates for the three biomarker searches. Conversely, the number of false positives was much lower for MWDA than for the *t*-tests. Finally, as is also clear from Figure 7, most true positives were already identified by the relatively simple direct NSI measurement of carbonic anhydrase. For MWDA, a number of positive *F* values clearly stick out and match with peaks in the mass spectrum. For the *t*-tests, some matches can also be observed, and the *t*-tests on permuted data do not show obvious matches, as expected. The LC/MS/MS measurements supply a few additional identifications.

DISCUSSION

With respect to the simulated data, the similarity of the performances of *t*-tests and MWDA for perfectly aligned data is not unexpected. A complete LC/MS measurement contains many correlated features, such as peptides originating from the same protein or peptides originating from proteins that are part of the same biochemical pathway. However, these peptides are far more likely to display different retention behaviors and to have different mass-to-charge ratios than identical ones. Therefore, apart from chance effects and near-perfect correlations within a peak or isotope pattern, a perfect (i.e., perfectly aligned, noiseless, etc.) LC/MS data set will show very few correlations in its individual mass spectra or chromatograms. As Zuber and Strimmer[30] pointed out, for noncorrelated data, LDA reduces to diagonal discriminant analysis (DDA), for which *t*-tests provide the optimally achievable results.[30,31]

When peaks are shifted, however, a *t*-test applied at the top of a peak in sample x will partly or completely miss the corresponding peak in sample y. Thus, the test "observes" a large variance (top and baseline), which ultimately obscures any remaining variance from differences between case and control samples. PCA−LDA, on the other hand, is able to combine variables that are more strongly correlated within the case and control groups, respectively, than between the two. Thus, as long as shifts are not overly large and the number of samples is not exceedingly small, shifted peaks can still be compared. For a large number of samples, it can be expected that the performance of the *t*-tests catches up with MWDA again, simply because of the $1/n^{1/2}$ relation in the test statistic.

As mentioned in the Theory section, the importance of the exact *p* or *F* values that are obtained for *t*-tests or MWDA, respectively, is limited. Their rank order is of greater importance here. Therefore, in retrospect, the significance levels that we used to separate LC matrices into positives and negatives were rather arbitrary. Another reasonable choice could have been to use a

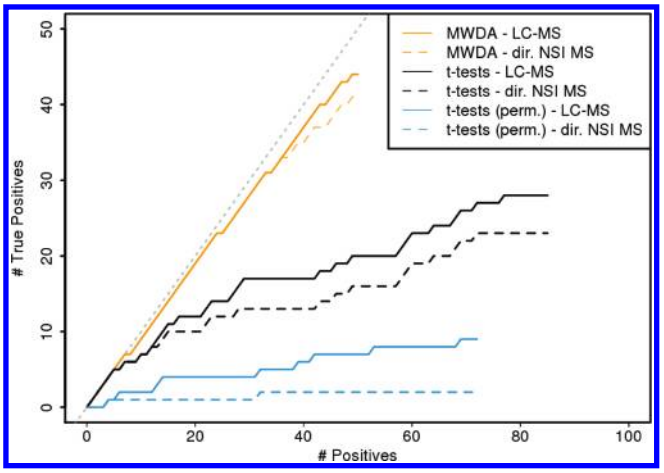


Figure 6. Partial true positive rates for MWDA, *t*-tests, and *t*-tests with permuted class labels applied to the *E. coli* set. Solid lines represent evaluation based on identifications by LC/MS; dashed lines are for evaluation based on direct NSI MS. The dotted gray line presents the maximally achievable true positive rate, that is, when all LC matrices designated positive are true positives.

predetermined cutoff for the number of positives, based on reasonable limits for the number of sequencing events in the final LC/MS/MS measurements used for identification purposes. All in all, the choice of a cutoff level should strike a balance between the work load associated with the identification and validation of larger numbers of positives and the need or wish to find more true positives by analyzing increasing numbers of positives with a decreasing likelihood of being true positives.

It is an interesting observation that MWDA selects multiple isotope peaks of the same pattern more often than *t*-tests. This increases the confidence that an actual biomarker has been found. One could argue about whether positives should be defined on the level of peaks, isotope patterns, peptides (including multiple charge states, adducts, and intramolecular rearrangements), or proteins (which is not really an alternative in this case). Here, we have chosen to present the positives as they are found, without an additional interpretation step.

Instead of using a fixed number of PCs in MWDA for all LC matrices, it might seem that optimizing the number of PCs per LC matrix, using a cross-validation approach for instance, might enhance the results. This is not the case, however! Generally, classification methods are applied to single matrices, and the hypothesis is that class information is contained in the matrix. Optimizing the number of PCs in that case is warranted. For the LC/MS data discussed here, however, the majority of the LC matrices are expected to contain *no* classification information. The effect of optimizing the classification performance on each and every matrix is that the positives get lost in the noise of negatives (results not shown) that are also optimized to give the best classification result possible. Therefore, choosing an overall number of PCs based on the discrimination between positive and negative matrices of the training set is actually the better approach. In practical situations, where no training set with a known ground truth is available, such as for the *E. coli* set, the optimal number of PCs needs to be chosen in a different, but still unbiased, way. Here, we adopted the frequently used approach of picking the number of PCs before the first elbow or knee in a scree plot.[11] The rounded average number of PCs for 10

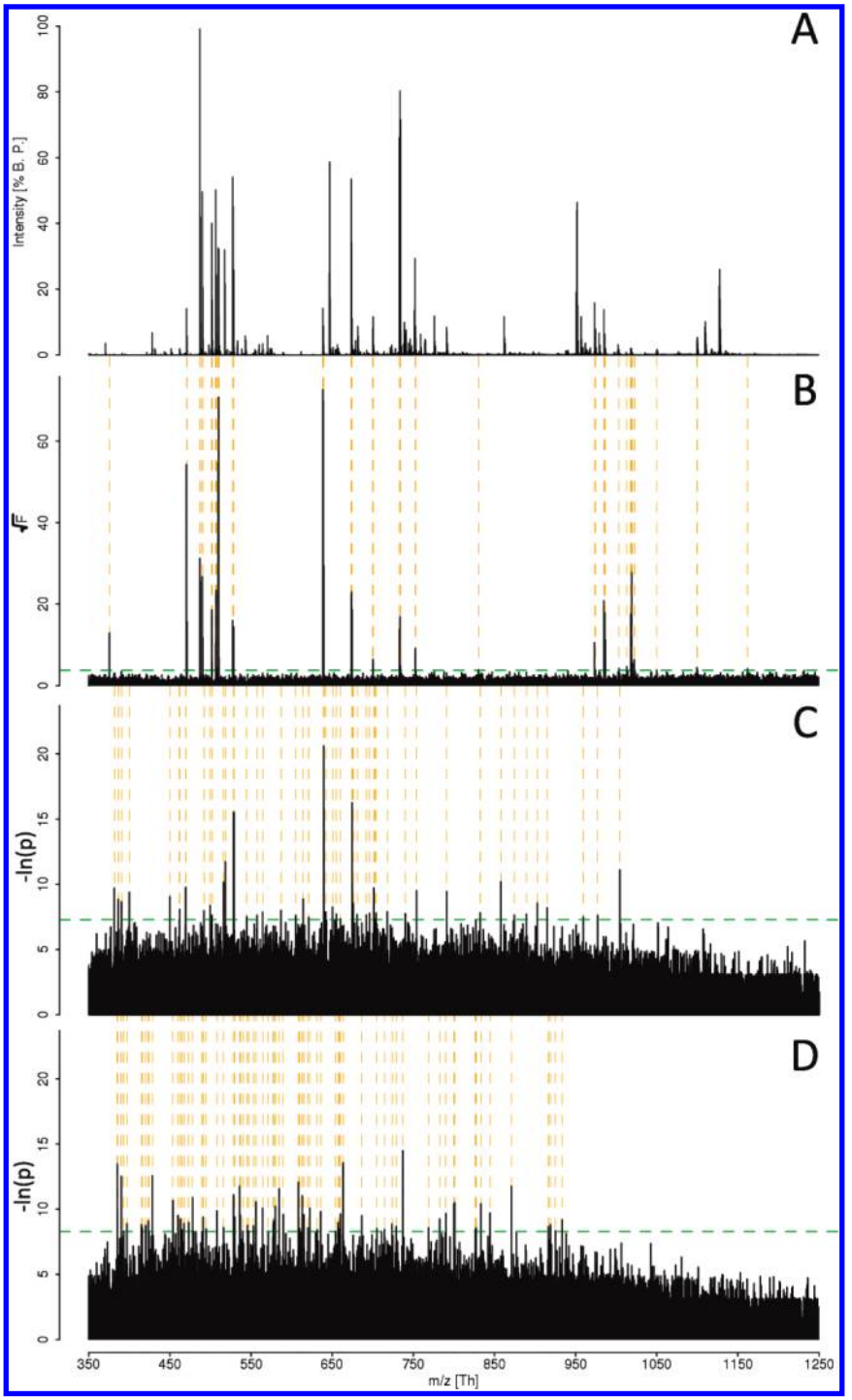


Figure 7. Direct comparison of (A) the carbonic anhydrase direct NSI mass spectrum, (B) the square roots of MWDA Fisher quotients, (C) $-\ln p$ for the *t*-tests, and (D) $-\ln p$ for the *t*-tests on permuted data, obtained for the *E. coli* data set. The green dashed horizontal lines in panels B and C are the respective 10% significance levels obtained from permutation tests. The similar line in panel D is the 4% cutoff used for the permuted data. Yellow dashed vertical lines were drawn as a guide to the eye from F/p values that are positives (i.e., above the significance level/cutoff).

randomly chosen LC matrices was used for all LC matrices. Other methods are conceivable, however.

In this work, we applied MWDA to only the LC matrices of LC/MS data sets. In principle, it is also possible to apply it to the MS matrices, in which case the F values form a chromatogram-like structure and the discriminant coordinates resemble mass spectra. However, in that case, misalignments larger than the chromatographic peak width along the retention time axis truly prohibit analysis: Because the shifts are orthogonal to the mass spectral direction, corresponding peaks will simply not be present in the same MS matrices anymore.

In principle, it is possible to correct for shifts by using alignment techniques.[32−34] When a good alignment is obtained, the results of both *t*-tests and MWDA will improve, with MWDA being robust

over a wider range of (remaining) misalignment. In our experience, obtaining a good alignment is not trivial, however; although alignment techniques will generally improve the agreement between retention times, in practice, perfect alignment is never achieved. Preliminary experiments for the *E. coli* data show that it is very well possible for biomarker search results to deteriorate upon use of data that have been aligned and are well-aligned according to generally used criteria such as the correlation between samples and visual inspection. Further investigations in this direction are under way.

Although we have focused on comparing a univariate and a multivariate technique for the analysis of LC/MS data, it should be mentioned that LC/MS data have an extra level of complexity; not only are they intrinsically multivariate, but they are also of *multiway* nature.[35,36] Multiway data need multiway data analysis, but the more sophisticated multiway methods such as parallel factor analysis (PARAFAC) and Tucker modeling impose very strict conditions on the data, which LC/MS data do not fulfill. We have, however, applied averaging and unfolding, two basic techniques from the field of multiway analysis that reduce the data cube to a single matrix and subsequently applied multivariate techniques in the normal fashion. The performances of these methods were considerably lower than those of both *t*-tests and MWDA, and therefore, we did not discuss them further here.

Finally, the MWDA method described here offers not only a solution for current bottom-up (peptide-based) proteomics but holds great promise for future top-down (protein-based) proteomics by LC/MS. Current advances in top-down instrumentation enable routine analysis of intact proteins by LC/MS. Electron-transfer dissociation (ETD) on these instruments allows for intact protein identification but requires manual optimization of MS/MS fragmentation settings for each individual protein. This criterion thus implies a limit to the number of proteins amenable to ETD analysis and, therefore, depends on methods such as MWDA that select ions of interest for further identification from MS-level data. We therefore believe that our current MWDA method holds great potential for application in top-down LC/MS-based proteomics in contrast to identification-driven quantitation strategies.

CONCLUSIONS

In this article, we have extended the use of a multivariate pattern recognition technique to complex LC/MS data. We applied the technique to a number of simulated data sets, as well as to real spike-in LC/MS data. We showed that the method is inherently more robust to misalignment imperfections than the commonly applied *t*-tests—resulting in significantly higher true positive rates—and that it identifies more biomarkers in a practical setting, characterized by a relatively low number of samples, misalignments, and noise. The current article is mainly a proof of principle, and more research to establish the exact application ranges of the method is warranted. However, given the robustness of the method to data imperfections, we believe the method to be of importance for the identification of biomarkers in current proteomics studies, as well as in upcoming top-down proteomics and other LC/MS analyses of complex samples (e.g., metabolomics).

The *E. coli* set discussed in the current article is part of a larger set of 59 samples that was measured according to a full factorial design of two factors (spikes) on three levels (concentrations). The full benchmark data set will be discussed in another work,[21] and the raw data (in mzXML format) will be made publicly available in an online repository. The binned 1801 × 2000 × 17 array used here will be made available in .mat (*Matlab*, The Mathworks, Inc., Natick, MA) and .RData (R[27]) formats.

ASSOCIATED CONTENT

Supporting Information. More elaborate descriptions of LDA and PCA−LDA, the simulated sets, sample preparation and data collection for the *E. coli* benchmark set and the carbonic anhydrase-only samples, and data processing steps prior to analysis. This material is available free of charge via the Internet at http://pubs.acs.org/.

AUTHOR INFORMATION

Corresponding Author

*E-mail: l.buydens@science.ru.nl (L.M.C.B.), ron.wehrens@iasma.it (R.W.). Phone: +31-24-3653180 (L.M.C.B.), +39-0461-615563 (R.W.). Fax: +31-24-3652653 (L.M.C.B.), +39-0461-650872 (R.W.).

Present Addresses

^Centro Ricerca e Innovazione, Fondazione Edmund Mach, Via E. Mach, 1, 38010, San Michele all'Adige (TN), Italy.

ACKNOWLEDGMENT

Part of this study was financially supported by the Princess Beatrix Foundation through Grant WAR05-0128. The authors thank Dr. J. T. M. Keltjens from the Department of Microbiology at Radboud University Nijmegen for kindly providing the *E. coli* culture.

REFERENCES

(1) Peng, J. M.; Gygi, S. P. *J. Mass Spectrom.* **2001**, *36*, 1083–1091.
(2) Ong, S.; Mann, M. *Nat. Chem. Biol.* **2005**, *1*, 252262.
(3) Hye, A.; Lynham, S.; Thambisetty, M.; Causevic, M.; Campbell, J.; Byers, H. L.; Hooper, C.; Rijsdijk, F.; Tabrizi, S. J.; Banner, S.; Shaw, C. E.; Foy, C.; Poppe, M.; Archer, N.; Hamilton, G.; Powell, J.; Brown, R. G.; Sham, P.; Ward, M.; Lovestone, S. *Brain* **2006**, *129*, 3042–3050.
(4) Etzioni, R.; Urban, N.; Ramsey, S.; McIntosh, M.; Schwartz, S.; Reid, B.; Radich, J.; Anderson, G.; Hartwell, L. *Nat. Rev. Cancer* **2003**, *3*, 243–252.
(5) Steen, H.; Mann, M. *Nat. Rev.* **2004**, *5*, 699–711.
(6) Marcotte, E. M. *Nat. Biotechnol.* **2007**, *25*, 755–757.
(7) Domon, B.; Aebersold, R. *Science* **2006**, *312*, 212–217.
(8) Wiener, M. C.; Sachs, J. R.; Deyanova, E. G.; Yates, N. A. *Anal. Chem.* **2004**, *76*, 6085–6096.
(9) Silva, J. C.; Denny, R.; Dorschel, C. A.; Gorenstein, M.; Kass, I. J.; Li, G.; McKenna, T.; Nold, M. J.; Richardson, K.; Young, P.; Geromanos, S. *Anal. Chem.* **2005**, *77*, 2187–2200.
(10) Listgarten, J.; Neal, R. M.; Roweis, S. T.; Wong, P.; Emili, A. *Bioinformatics* 2007, *23*, E198–E204.
(11) Jolliffe, I. T. *Principal Component Analysis*, 2nd ed.; Springer Series in Statistics; Springer: New York, 2002.
(12) Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; De Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics*; ; Elsevier: New York, 1997; Vol. 20A
(13) Vandeginste, B. G. M.; Massart, D. L.; Buydens, L. M. C.; De Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics*; ; Elsevier: New York, 1998; Vol. 20B
(14) Fisher, R. A. *Ann. Eugen.* **1936**, *7*, 179–188.
(15) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 1st ed.; Springer, 2001.
(16) Ami, D.; Natalello, A.; Mereghetti, P.; Neri, T.; Zanoni, M.; Monti, M.; Doglia, S. M.; Redi, C. A. *Spectrosc. Int. J.* **2010**, *24*, 89–97.

(17) Charlton, A.; Allnutt, T.; Holmes, S.; Chisholm, J.; Bean, S.; Ellis, N.; Mullineaux, P.; Oehlschlager, S. *Plant Biotechnol. J.* **2004**, *2*, 27–35.
(18) Cozzolino, D.; Smyth, H. E.; Cynkar, W.; Dambergs, R. G.; Gishen, M. *Talanta* **2005**, *68*, 382–387.
(19) Kher, A.; Mulholland, M.; Green, E.; Reedy, B. *Vib. Spectrosc.* **2006**, *40*, 270–277.
(20) Mertens, B. J. A. *J. Proteom.* **2009**, *72*, 785–790.
(21) Wessels, H. J. C. T.; Bloemberg, T. G.; Van Dael, M.; Wehrens, R.; Buydens, L. M. C.; Van den Heuvel, L. P.; Gloerich, J., manuscript submitted.
(22) Kelleher, N. L. *Anal. Chem.* **2004**, *76*, 196 A–203 A.
(23) B., B.; Smith, R. D. *Mass Spectrom. Rev.* **2005**, *24*, 168–200.
(24) Miller, J. N.; Miller, J. C. *Statistics and Chemometrics for Analytical Chemistry*; Pearson Education Ltd.: Harlow, Essex, U.K., 2005.
(25) Ballabio, D.; Skov, T.; Leardi, R.; Bro, R. *J. Chemom.* **2008**, *22*, 457–463.
(26) Zou, H.; Hastie, T. *J. R. Stat. Soc. B: Stat. Methodol.* **2005**, *67*, 301–320.
(27) R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2009.
(28) Wessels, H. J. C. T.; Gloerich, J.; der Biezen, E.; Jetten, M. S.; Kartal, B. *Methods Enzymol.* **2011**, *486*, 465–482.
(29) Fawcett, T. *Pattern Recogn. Lett.* **2006**, *27*, 861–874.
(30) Zuber, V.; Strimmer, K. *Bioinformatics* **2009**, *25*, 2700–2707.
(31) Fan, J.; Fan, Y. *Ann. Stat.* **2008**, *36*, 2605–2637.
(32) Bloemberg, T. G.; Giskeødegård, G. F.; Postma, G. J.; Sitter, B.; Tessem, M. B.; Gribbestad, I. S.; Bathen, T. F.; Buydens, L. M. C. *Anal. Chim. Acta* **2010**, *683*, 1–11.
(33) Bloemberg, T. G.; Gerretzen, J.; Wouters, H. J. P.; Gloerich, J.; Van Dael, M.; Wessels, H. J. C. T.; Van den Heuvel, L. P.; Eilers, P. H. C.; Buydens, L. M. C.; Wehrens, R. *Chemom. Intell. Lab. Syst.* **2010**, *104*, 65–74.
(34) Jellema, R. H. *Comprehensive Chemometrics*; ; Elsevier B. V.: Amsterdam, 2009; Vol. 2, Chapter Variable Shift and Alignment, pp 85−108.
(35) Kroonenberg, P. M. *Applied Multiway Data Analysis*; Wiley Series in Probability and Statistics; John Wiley & Sons, Inc.: New York, 2008.
(36) Smilde, A. K.; Bro, R.; Geladi, P. *Multi-way Analysis: Applications in the Chemical Sciences*; John Wiley & Sons, Ltd.: Chichester, U.K., 2004.