Automatic Deconvolution of Isotope-Resolved Mass Spectra Using Variable Selection and Quantized Peptide Mass Distribution

ARTICLE *in* ANALYTICAL CHEMISTRY · JUNE 2006

Impact Factor: 5.64 · DOI: 10.1021/ac052212q · Source: PubMed

Ruth Hogue Angeletti

Albert Einstein College of Medicine

Automatic Deconvolution of Isotope-Resolved Mass Spectra Using Variable Selection and Quantized Peptide Mass Distribution

Peicheng Du* and Ruth Hogue Angeletti

Department of Developmental and Molecular Biology, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, New York 10461

We present an algorithm for the deconvolution of isotope-resolved mass spectra of complex peptide mixtures where peaks and isotope series often overlap. The algorithm formulates the problem of mass spectrum deconvolution as a classical statistical problem of variable selection, which aims to interpret the spectrum with the least number of peptides. The LASSO method is used to perform automatic variable selection. The algorithm also makes use of the quantized distribution of peptide masses in the NCBInr database after in silico trypsin digestion as filters to aid the deconvolution process. Errors in the expected isotope pattern are accounted for to avoid spurious isotope series. The effectiveness of the algorithm is demonstrated with annotated ESI spectrum of known peptides for which the peaks and isotope series are highly overlapping. The algorithm successfully finds all correct masses in the experimental spectrum, except for one spectrum where an additional refinement procedure is required to obtain the correct results. Our results compare favorably to those from a widely used commercial program.

Deconvolution of the mass spectra of complex peptide mixtures is an increasingly important problem. Mass spectrometry (MS) has been widely used for protein identification, protein expression profiling, identifying posttranslational modifications, and studying protein structure and dynamics, among others. With the increased accessibility of mass spectrometers to researchers, mass spectrum data have been rapidly accumulating. One of the fundamental problems in protein mass spectrometry data processing is to automatically decompose the raw spectrum to a list of peptide masses or peptide product masses in the case of tandem MS. An automatic spectrum deconvolution method should take into account the isotope patterns of peptides, the presence of multiple charge states for electrospray ionization (ESI) spectra, and noise. Due to the natural abundance of heavy isotopes, particularly ^{13}C, several peaks exist for each peptide mass at a given charge state. These peaks are referred to as the isotope series. The isotope pattern refers to the isotope envelope or relative intensity ratios of the peaks in the same isotope series. The isotope pattern of a peptide is determined by the elemental formula of the peptide and the abundance of heavy isotopes, usually the natural abundance, and therefore known. Due to the presence of isotope series and multiple charge states for each peptide, and the fact that peptide masses tend to be clustered around certain values because elements CNOSH all have near-integral masses,[1,2] peaks and isotope series in a complex spectrum often overlap. With the interference of both instrumental and chemical noise, deconvolution can be highly nontrivial for multicomponent spectra.

Previous methods have been proposed for deconvolution of isotope-resolved MS spectrum. An incomplete list of such work includes Senko's method,[3] ZSCORE,[4] THRASH,[5] ISOCONV,[6] and ESI-ISOCONV,[7] Wang's method,[8] quadratic deisotoping,[9] Matching,[10] PepList,[11] and the program by Zhang.[12] With the exception of ZSCORE, all of these methods compare the observed isotope pattern to the expected isotope pattern. Senko's method and ZSCORE fail when peaks overlap. MATCHING can perform deconvolution of overlapping isotope series only if the peptide sequences and therefore masses are known a priori. The quadratic deisotoping method treats the intensities of isotope series as unknowns and tries to calculate them by solving a quadratic programming problem. Other methods generally take a stepwise approach; i.e., first find an isotope series that fits well to the expected isotope pattern, then subtract it from the spectrum, and iterate until no more isotope series can be found. These methods are insufficient for complex spectra with overlapping isotope series

* To whom correspondence should be addressed. E-mail: pdu@aecom.yu.edu.

(1) Mann, M., *43rd ASMS Conference on Mass Spectrometry and Allied Topics*, Atlanta, GA, 1995; p 639.
(2) Gay, S.; Binz, P. A.; Hochstrasser, D. F.; Appel, R. D. *Electrophoresis* **1999**, *20*, 3527−3534.
(3) Senko, M. W.; Beu, S. C.; McLafferty, F. W. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 229−233.
(4) Zhang, Z.; Marshall, A. G. *J. Am. Soc. Mass Spectrom.* **1998**, *9* (3), 225−33.
(5) Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. *J. Am. Soc. Mass Spectrom.* **2000**, *11* (4), 320−32.
(6) Wehofsky, M.; Hoffman, R.; Hubert, M.; Spengler, B. *Eur. J. Mass Spectrom.* **2001**, *7*, 39−46
(7) Wehofsky, M.; Hoffman, R. *J. Mass Spectrom.* **2002**, *37*, 223−229.
(8) Wang, W.; Zhou, H.; Lin, H.; Roy, S.; Shaler, T. A.; Hill, L. R.; Norton, S.; Kumar, P.; Anderle, M.; Becker, C. H. *Anal. Chem.* **2003***, 75*, 4818−4826
(9) Samuelsson, J.; Dalevi, D.; Levander, F.; Rögnvaldsson, T. *Bioinformatics* **2004**, *20*, 3628−3635.
(10) Fernández-de-Cossio, J.; Gonzalez L. J.; Satomi Y.; Betancout L.; Ramos Y.; Huerta V.; Besada, V.; Padron, G.; Minamino, N.; Takao, T. *Rapid Commun. Mass Spectrom.* **2004**, *18*, 2465−2472.
(11) Li, X.; Yi, E. C.; Kemp, C. J.; Zhang, H.; Aebersold, R. *Mol. Cell. Proteomics* **2005***, 4*, 1328−1340.
(12) Zhang X.; Hines, W.; Adamec, J.; Asara, J. M.; Naylor, S. *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 1181−1191.

10.1021/ac052212q CCC: $33.50 © 2006 American Chemical Society
Published on Web 04/11/2006

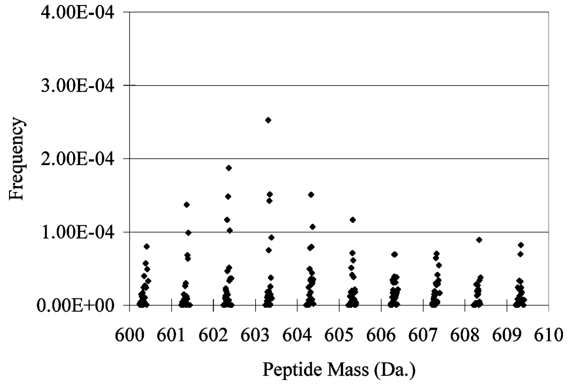

Figure 1. Histogram of quantized peptide mass distribution from 600 to 610 Da in NCBInr after virtual trypsin digestion, allowing three miss cleavages and no modification. Each diamond represents the frequency of occurrence of a mass bin with bin width of 0.01 Da. Clearly some mass values do not exist.

or peaks, for which there can be multiple potential interpretations (or models) to explain the data. In addition, the expected isotope patterns have errors that need to be accounted for. The methods mentioned above are neither systematic nor optimal in terms of evaluating all potential models and giving the best interpretation. Neither is some commercial software, such as BioAnalyst (AB, Foster City, CA).

The deconvolution problem can be formulated as a problem of variable selection for which many statistical theories and methods exist.[13] Variable selection refers to the problem of selecting the most relevant subset among multiple variables for the purpose of interpretation or prediction. Since a mass spectrum is essentially the intensity-weighted sum of multiple isotope series, deconvolution is a problem of finding the intensity of all the possible isotope series. An isotope series exists on the spectrum only if its intensity is positive. If we consider the intensities of isotope series as unknown variables, deconvolution becomes a variable selection problem where the goal is to select the most relevant subset of variables to explain the spectrum. A variable selection approach of deconvolution aims at selecting the simplest model, i.e., the model with the least number of isotope series. For a given spectrum, the variable selection procedure considers all possible isotope series and charge states, decides which ones exist, and calculates their intensities. Since it considers all possible explanations of the spectrum, the variable selection method of deconvolution is a systematic solution based on well-tested statistical theories.

This algorithm is novel not only because it uses the variable selection approach but also in two other aspects. First, it makes use of the quantized distribution of peptide masses in the NCBInr database after in silico trypsin digestion to aid deconvolution (Figure 1). The quantized peptide mass distribution can be used as a filter since some peptide mass values are nonexistent. This helps deconvolution by reducing the number of potential isotope series considered. Second, it explicitly accounts for errors in the expected isotope pattern by using the importance value (explained in the Methods section).

To assess the effectiveness of this algorithm, two complex ESI-MS spectra are used. The first one is the spectrum of a known peptide mixture with eight unique masses, each of which is 1 Da apart from another mass; thus their isotope series overlap. The results are compared to those obtained from BioAnalyst for the same spectrum. To further test the method, a more complex spectrum of rat serum is used.

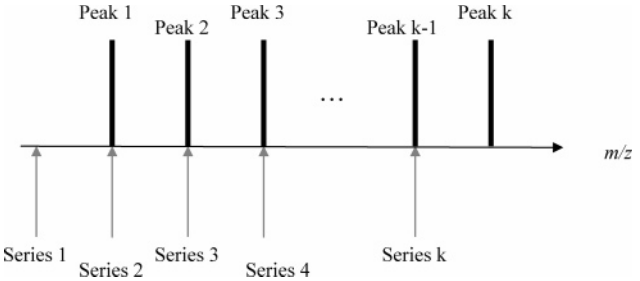

Figure 2. Building a master list from a list of k peaks with the spacing between consecutive peaks of $1/z$. Each vertical arrow points to the monoisotopic peak location of an isotope series of charge z. The master list contains all possible isotope series and charge states.

METHODS

Baseline Subtraction, Peak Detection, and Peak List Partitioning. The first step in deconvolution is to subtract the baseline because a nonflat baseline can distort the isotope pattern. For a spectrum with well-resolved peaks, a simple and effective way to calculate the baseline is to use the minimum of a sliding 4 m/z window along the m/z axis as the baseline. For a spectrum with not so well-resolved peaks, the size of the sliding window can be increased to include at least one m/z channel without signal. After the baseline is subtracted, the spectrum is reduced to a peak list with a suitable peak detection algorithm, such as searching for local maximum points above a preset threshold. Subsequently, the peak list is divided into local peak lists such that two consecutive peaks are separated into two local peak lists if and only if the spacing between them is over 2 m/z. The 2 m/z spacing ensures that peaks which could possibly belong to the same isotope series are put into the same local peak list. Then variable selection is applied to each local peak list sequentially from low m/z to high m/z. The order does not matter since each local peak list is treated independently in this step.

Building the Master List with Quantized Mass Distribution as a Filter. The next step is to build a master list, which includes all the possible isotope series for each local peak list. Each element on the list is an isotope series of a charge state, from which a charge zero monoisotopic mass can be calculated by assuming the adduct ion is H^+. The expected isotope pattern for each mass is represented by the averagine model.[3] The master list is essentially a candidate list of monoisotopic masses to be considered, except that more than one element on the list may correspond to the same mass but different charge states. Assuming the highest charge state that is isotope-resolved is "c" and an isotope series has at least two peaks, a master list can be derived from a local peak list as follows: for each charge state z, where $z = 1, 2, ..., c$, find the longest sequence of k peaks ($k \geq 2$) where the spacing between consecutive peaks in the sequence is $1/z$ within a preset tolerance value (Figure 2), and then consider each of the first $k - 1$ peaks of the peak sequence as the monoisotopic peak of a charge z isotope series and add them to the master list. The isotope series with a monoisotopic peak at $1/z$ to the left of

[13] George, E.I. *J. Am. Stat. Assoc.* **2000**, *95*, 1304–1308.

the first peak, such as series 1 in Figure 2, is also considered, because it has at least two peaks. Next, if any of the two most intense peaks in the averagine model is missing for an isotope series, the series is removed from the master list. Note that peaks can be reconsidered for as many charge states as possible. Eventually, the master list is constructed without using the peak intensity information. And the list has all applicable isotope series represented by their monoisotopic masses and their corresponding charge states.

Since the peptide mass distribution is discrete, this distribution can be used to filter the master list to remove isotope series that do not actually exist to reduce the complexity of the model. For example, for a peak list of four peaks in the order of 601.8, 602.3, 602.8, and 603.3 *m/z*, regardless of peak intensity, three doubly charged isotope series can be added to the master list, each starting from 601.8, 602.3, and 602.8, respectively, together with two singly charged isotope series starting from 601.8 and 602.3, respectively. However, the singly charged series starting from 601.8 corresponds to a monoisotopic mass of 600.8 (assuming the adduct ion is H^+), which has a frequency of occurrence of zero according to the mass distribution in Figure 1. Thus, that isotope series does not exist and is removed from the master list. It should be noted that the quantized mass distribution as a filter should be used with caution for nonpeptide peaks.

Using Variable Selection To Select Isotope Series That Exist. Next, since a mass spectrum is essentially the intensity-weighted sum of multiple isotope series, we can express the intensity of each peak on the local peak list as a weighted sum of contributions from each isotope series on the master list, with each contribution being the fraction of the corresponding peak in the averagine model of that series. The variable selection procedure selects the smallest subset of isotope series on the master list that explains the spectrum well. Variable selection is performed by the LASSO method.[14] For a total of *N* peaks and *P* isotope series on the master list, LASSO finds the intensity of each isotope series to minimize the sum of square residuals of the following system of linear equations:

$$y_i = \sum_{j=1}^{p} x_j \, \mathrm{f}_{jk} (1+w_j) \tag{1}$$

Subject to the constraint

$$\sum_{j=1}^{P} |x_j| \le s$$

Where y_i is the peak intensity of the *i*th peak, $i = 1, 2, ...N$; x_j is the unknown intensity of the *j*th isotope series on the master list, $j = 1, 2, ...P$; f_{jk} is the intensity of the *k*th isotopic peak in the averagine model of the *j*th isotope series, such that the *k*th isotopic peak and the *i*th peak have the same *m/z*, $k \le 5$; w_j is a small weight factor between 0 and 0.2, which is assigned to the *j*th isotope series such that isotope series of higher charge states are assigned a slightly bigger weight (see Supporting Information for the calculation of w_j); *s* is set to 80% of the total peak intensity (also see Supporting Information for the derivation). It is a tradeoff between false positives and false negatives to choose *s*. If *s* is too big, there would be too many false positives. If *s* is too small, the number of false negatives would increase. Thus, the value of *s* we choose reflects the accuracy of the averagine model, and it also does well with the test spectra.

Additional Filtering and Pooling of Results. LASSO returns as results a list of isotope series intensities with some being zero. We apply additional filters to the isotope series with nonzero intensities. Because the averagine is only an approximate isotope pattern and usually has ~20% error, weak isotope series could be consequences of fitting to that error. To avoid such weak series, which are mostly spurious, we calculate the value "importance" to measure the relative contribution of the isotope series to explain data as follows:

$$\text{importance} = \sum_{k=1}^{5} (xf_k^2)/y_k' \tag{2}$$

where f_k is the intensity of the *k*th peak in the averagine model; *x* is the calculated intensity for that isotope series by LASSO; y_k' is the fitted peak intensity of the peak that matches the *k*th peak of that isotope series. Note that y_k' is different from y_k, the actual peak intensity of that peak. Equation 2 can be interpreted as follows: Since xf_k is the peak intensity contributed by the *k*th isotopic peak, xf_k/y_k' is the relative contribution of that isotope series to that matching peak. Since the relative contribution by a strong isotopic peak means more than that of a weak isotopic peak, we weight the relative contributions of each isotopic peak by f_k, which gives us xf_k^2/y_k'. Summing up xf_k^2/y_k' for all isotopic peaks gives us the right side of eq 2.

Since the importance value measures the relative contribution of an isotope series in explaining the data, we disregard the isotope series for which the importance value is below 1/3, which is about the upper limit of the error range of the averagine model, to be conservative. The intensity of an isotope series is set to zero once it is removed from consideration. With the isotope series for which the intensity still remains positive, we reconstruct eq 1 with the nonzero x_j without the weight w_j and solve the system of equations as a usual least-squares regression to determine the intensity values and their estimated standard errors.

Next, the list of isotope series with positive intensities from all local peak lists are combined to check for different charge states of the same mass. The final intensity of each mass is the sum of intensities at all charge states of that mass. To make use of multiple peaks for the same mass to improve mass accuracy, we recalculate the monoisotopic mass as an intensity-weighted average from the *m/z* of several isotopic peaks as in THRASH.[5]

Assessing the Performance of the Deconvolution Algorithm. The algorithm is tested with the ESI-MS spectrum of a mixture of 16 known synthetic peptides with 8 unique masses (Figure 3a); i.e., for each unique mass there are two peptides. The monoisotopic masses of them are as follows: 947.55, 948.54, 962.53, 963.51, 1023.58, 1024.57, 1038.56, and 1039.54, respectively, with sequences of VFLQSLKN, VFLQSIKN, VFLQSLKD, VFLQSIKD, NFLQSLKD, NFLQSIKD, NFLQSLKN, NFLQSIKN, VFLQYLKN, VFLQYIKN, VFLQYLKD, VFLQYIKD, NFLQYLKN, NFLQYIKN, NFLQYLKD, and NFLQYIKD.

[14] Tibshirani, R. *J. R. Stat. Soc. B* **1996**, *58* (1), 267−288.

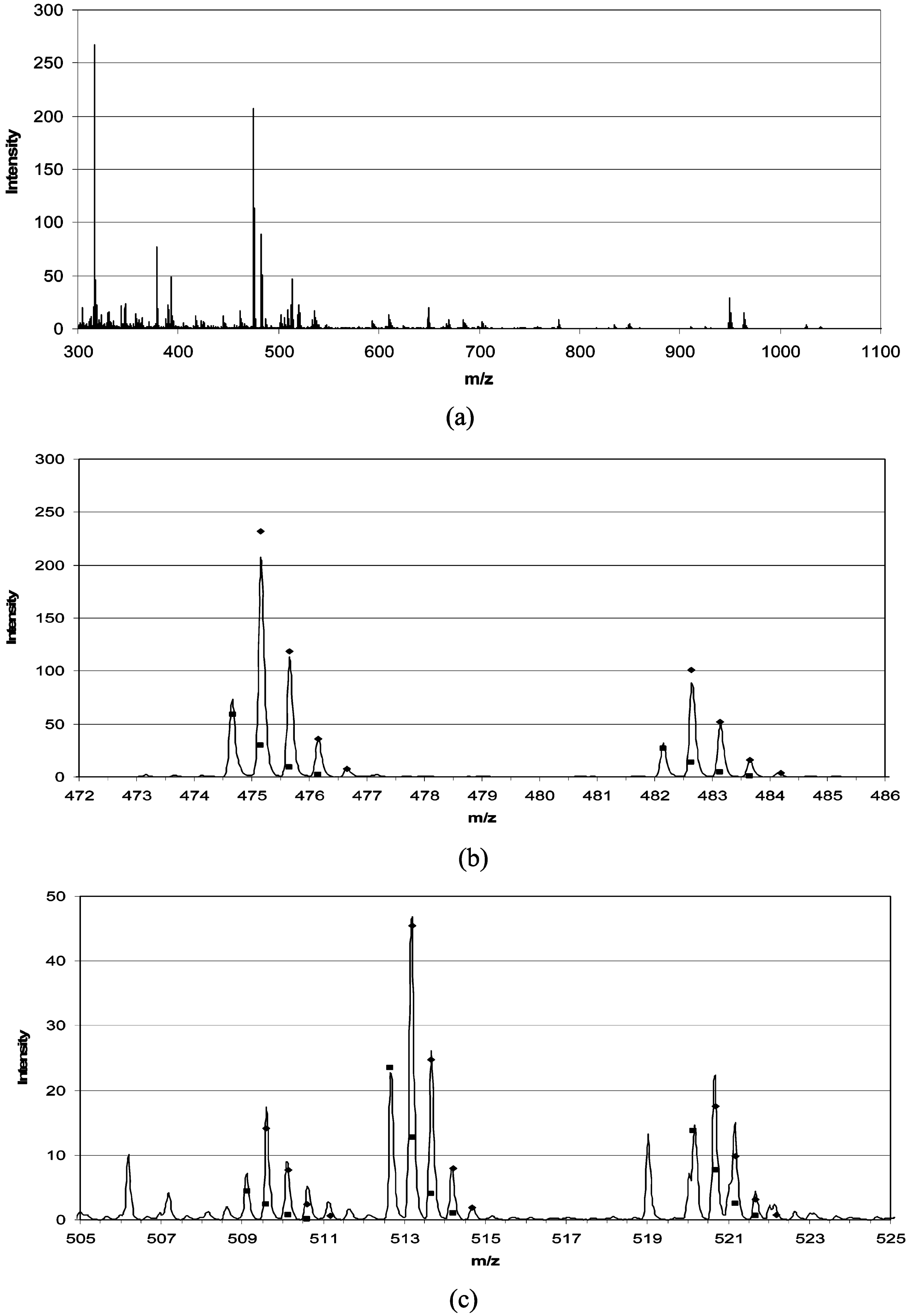

Figure 3. ESI spectrum of a mixture of 16 known peptides with 8 unique masses. (a) The complete spectrum, (b) The region of 472−486 *m*/*z*. (c) The region of 505−525 *m*/*z*. Filled squares in (b) and (c) represent contributions from the lower mass peptides in overlapping isotope series. Filled diamonds represent contributions from the higher mass peptides. These contributions are calculated by linear fitting with averagine; therefore, they do not exactly match the actual peak intensities.

The experimental conditions for the mixture of 16 peptides are described as follows. Samples were analyzed by a Qq-TOF mass spectrometer (Qstar Pulsar, AB, Foster City, CA). The 2% CH_3CN/0.1% formic acid was used as solvent A. The 80% CH_3CN /0.1% formic acid was used as solvent B. A C18 column with 300 μm i.d. × 15 cm (Dionex, CA) was used for the separation. The following gradient was used for LC: 30 min at 5% solvent B (desalting) followed by 5−55% B from 30 to 80 min. The flow rate was 3 μL/min. Microelectrospray sources with 20-μm-i.d. capillary was used for ESI. TOF-MS scan was performed in the *m/z* 300−1800 with a scan time of 1 s. The instrument was calibrated using CsI−peptide mixture as suggested by the vendor. The 16-peptide mixture containing 2.5×10^{-9} M concentration of each peptide in a 50-μL volume was injected. The spectrum is the averaged spectrum from 55.0 to 76.5 min. Deconvolution with BioAnalyst is performed with Bayesian Peptide Reconstruct using the default parameters.

The rat serum spectrum was obtained with the same mass spectrometer. Normal serum was trypsin digested prior to injection. The C18 column used is 75 μm i.d. × 25 cm (Dionex). The gradient and solvents for LC were the same, except the elution time was 30−90 min. The flow rate was 250 nL/min. Nanoelectrospray sources were used instead of microelectrospray without capillary. For peaks in Figure 4 annotated with MS/MS data by MASCOT (Matrix Science, London, UK), the *m/z* of the monoisotopic peaks and their sequences are as follows: 587.34, LLWESGSLLR; 616.33, VFSPNVLNWR; 646.79, ITDNMFCAGFK; 661.35, IAELFSDLEER or IAELFSELDER; 825.91, TDVTQQLNTLFQDK.

RESULTS AND DISCUSSION

Example 1: Spectrum of a Mixture of Eight Unique Masses. The algorithm is first tested with the ESI-MS spectrum of a mixture of 16 known peptides with 8 unique masses: 947.55, 948.54, 962.53, 963.51, 1023.58, 1024.57, 1038.56, and 1039.54 Da. These peptides are known to generate doubly charged ions under the experimental conditions used. If the algorithm works correctly, it should find these eight masses even though they severely overlap with each other. At the same time it should not generate false positives. The spectrum is shown in Figure 3a. As can be seen, there are some impurities which we ignore unless they interfere with these peptides.

In the region of 474−477 *m/z* where doubly charged ions of monoisotopic masses 947.55 and 948.54 are expected (Figure 3b), five peaks are present: 474.68, 475.15, 475.65, 476.16, and 476.66 *m/z*. On the master list, there are four doubly charged isotope series starting from 474.68, 475.15, 475.65, and 476.16 *m/z*, respectively, and three singly charged series starting from 474.68, 475.15, and 475.65 *m/z*, respectively. The singly charged series starting from 474.68 and 475.65 *m/z* are later removed from the master list because there are no corresponding monoisotopic masses within 0.2 Da. according to the quantized peptide mass distribution. LASSO reports that only the two doubly charged series corresponding to monoisotopic masses of 947.32 and 948.29 have nonzero intensities, which match masses 947.55 and 948.54, respectively, within experimental error. Even without using quantized mass distribution as a filter, LASSO still reports the same results.

Likewise, in the region of 482−486 *m/z* where doubly charged ions of monoisotopic masses 962.53 and 963.51 are expected (Figure 3b), LASSO reports that only the two doubly charged series corresponding to monoisotopic masses of 962.28 and 963.27 exist, matching masses 962.53 and 963.51, respectively, within experimental error. In the region of 505−515 *m/z* where doubly charged ions of 1023.58 and 1024.57 *m/z* are expected (Figure 3c), in addition to the two masses 1023.33 and 1024.34, LASSO also reports a singly charged isotope series corresponding to a monoisotopic mass of 505.19, and two doubly charged isotope series corresponding to monoisotopic masses of 1016.21 and 1017.18. The three extra masses found are probably either solvent ions or impurities.

In the region of 518−524 *m/z* where doubly charged ions of monoisotopic masses 1038.56 and 1039.54 are expected (Figure 3c), LASSO reports two doubly charged series corresponding to monoisotopic masses 1038.31 and 1039.33, matching masses 1038.56 and 1039.54, respectively, within experimental error. In addition, LASSO also reports two extra monoisotopic masses 518.04 and 520.15, both singly charged. The monoisotopic mass 520.15 is later removed because its importance value is only 0.3, less than the 1/3 threshold value. It is possible that this mass actually exists. However, it is removed because the algorithm considers it more likely an artifact due to errors in the averagine model than an actual mass. The use of the importance value as a filter may cause a real peptide to be filtered out, if its intensity is less than 1/10 of the overlapping peptides, and its strongest isotopic peak overlaps with those of stronger peptides. This is a consequence of the inaccuracy of the averagine model. Another monoisotopic mass 518.04 clearly exists on the spectrum, which is probably a peptide. In this region of 518−524 *m/z*, the algorithm successfully separates three overlapping isotope series of different charge states.

In comparison to BioAnalyst, among the eight known masses, BioAnalyst only finds four in the same regions of the spectrum in Figure 3b and Figure 3c: masses 948.31, 963.29, 1024.34, and 1039.32, matching 948.54, 963.51, 1024.57, and 1039.54 Da, respectively.

Example 2: Partially Annotated Spectrum of Rat Serum. It is a single scan spectrum at 78.0 min, chosen because some peaks on the spectrum have been identified with MS/MS. For straightforward cases without overlapping peaks or isotope series, the algorithm correctly finds the peptide masses. To evaluate the performance of the algorithm on overlapping peaks or isotope series, five segments of the spectrum are examined in detail (Figure 4). The first spectrum segment is 585−591 *m/z* (top left). It is known to have a doubly charged isotope series starting from 587.34 *m/z*, and a singly charged isotope series starting from 588.34 *m/z*, according to both protein identification results of the peptide at 587.34 *m/z* with MS/MS and its spectrum at 76.7 min (top right). The second spectrum segment is 614−620 *m/z* (second row, left). This spectrum is known to have two doubly charged isotope series starting from 616.34 and 616.82 *m/z*, respectively, according to both the protein identification results of the peptide at 616.34 *m/z* with MS/MS and its spectrum at 78.4 min (second row, right). The third spectrum segment is 644−650 *m/z* (third row, left). This segment is known to be a mixture of two doubly charged isotope series, each starting from 645.32

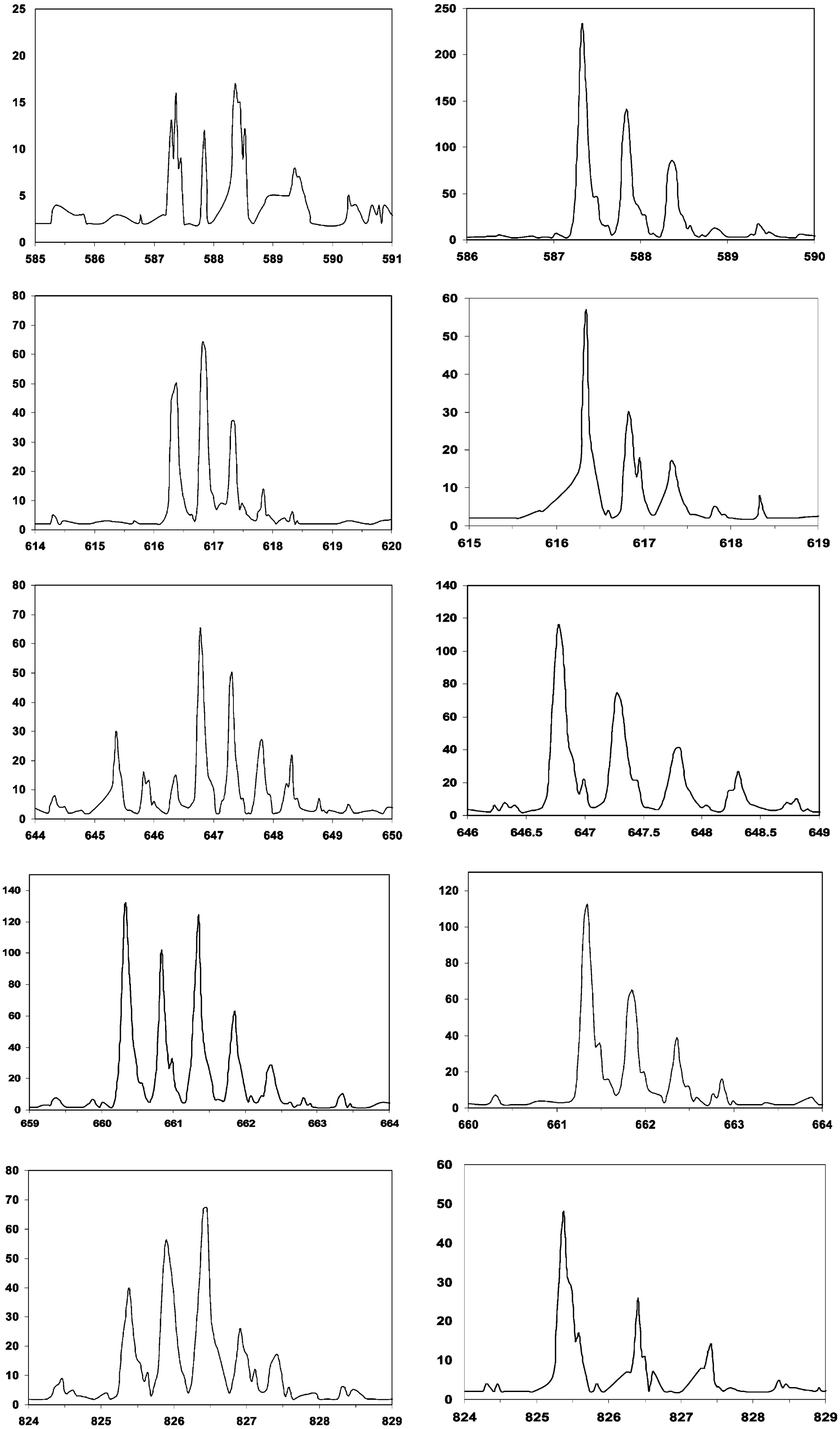

Figure 4. ESI spectrum of a serum fraction. The *x* axis is *m*/*z* and the *y* axis is intensity. Top row: (left) the spectrum segment at 78.0 min; (right) the spectrum segment at 76.7 min. Second row: (left) the spectrum segment at 78.0 min; (right) the spectrum segment at 78.4 min; Third row: (left) the spectrum segment at 78.0 min; (right) the spectrum segment at 77.5 min; Fourth row: (left) the spectrum segment at 79.0 min; (right) the spectrum segment at 78.0 min; Bottom row: (left) the spectrum segment at 78.0 min; (right) the spectrum segment at 77.3 min.

and 646.79 *m/z*, respectively, based on both the protein identification results of the peptide at 646.79 *m/z* with MS/MS and its spectrum at 77.5 min (third row, right). The fourth spectrum segment is 659−664 *m/z* (fourth row, left) at 79.0 min. This segment is known to have two doubly charged isotope series, each starting from 660.34 and 661.34 *m/z*, respectively, according to both the protein identification results of the peptide at 661.34 *m/z* with MS/MS and its spectrum at 78.0 min (fourth row, right). For all four cases mentioned above, the algorithm successfully finds the correct peptides without any false positives in these spectrum segments. The intensities of the peptides found seem to be reasonable as well.

The algorithm is not foolproof (as is the case with any other algorithm). An example of this is the fifth spectrum segment at 824−829 *m/z* (Figure 4, bottom left). This spectrum is annotated to be a mixture of a doubly charged isotope series starting from 825.91 *m/z* and a singly charged isotope series starting from 825.38 *m/z*, according to protein identification results of the peptide at 825.91 *m/z* with MS/MS and the spectrum of the singly charged series at 77.3 min (Figure 4, bottom right). The algorithm interprets this spectrum to have a singly charged isotope series at 826.40 *m/z*, and two doubly charged series starting from 825.38 and 825.91 *m/z*, respectively. Therefore, the algorithm interprets the spectrum as having three peptides when it can be explained equally well by the two correct peptides, which suggests a failure in the search process to find the least number of peptides to explain the spectrum. Such failure is expected because of two main reasons. First, the algorithm has to use approximations, such as the averagine model. Second, there may be more than one interpretation that explains the spectrum almost equally well.

To overcome the inaccuracy of the averagine model and check the existence of multiple interpretations that are equally good, a refinement procedure is developed. Refinement starts from the LASSO results. Let P denote the number of peptides reported by LASSO. For each possible combination of P peptides, the minimum sums of square errors (SSE) are calculated by allowing the number of carbons (thus the isotope pattern) for each mass to vary within a normal range (found in the sequence database) to minimize the SSE, instead of using the averagine model. The minimum SSE for each combination of P peptides is used as the SSE for that peptide combination to represent the goodness of fit. Next, if $P - 1 > 0$ or $P - 2 > 0$, all possible combinations of $P - 1$ or $P - 2$ peptides are evaluated for goodness of fit as well. For the bottom left spectrum in Figure 4 for which LASSO gives incorrect results, the square root of SSE for all peptide combinations with $P - 2$, $P - 1$, or P peptides are plotted in Figure 5 where $P = 3$. The overall trend is the more the peptides, the better the fit, which is expected because more variables generally allow better fitting. The combination with the two correct peptides has the lowest SSE among all two-peptide combinations. Though there are six three-peptide combinations, including the one found by LASSO, with similar or lower SSE than that of the correct two-peptide combination, the small difference in SSE does not justify one more variable, because an extra variable may slightly improve the SSE just by chance. Therefore, by visual inspection, the correct two-peptide combination is the best interpretation according to the principal of variable selection. For a quantitative model selection criteria, a commonly used one in the statistics literature is adjusted R^2 (Supporting Information, Appendix 2).

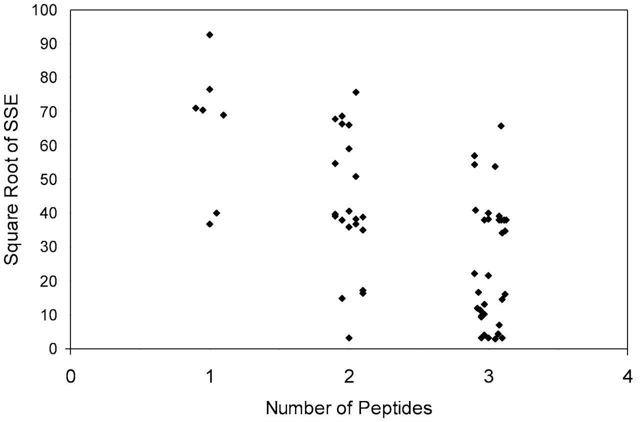

Figure 5. Distribution of SSE vs the number of peptides used to fit the spectrum at bottom left of Figure 4. The *x* axis represents the number of peptides, and the *x* values are scattered around the integers to avoid point overlapping. The correct peptide combination is the lowest point among two peptide combinations. The peptide combinations are drawn from seven peptides on the master list, of which three are singly charged, starting from 825.38, 825.89, and 826.41, respectively; four are doubly charged, starting from 825.38, 825.89, 826.41, and 826.92, respectively.

The refinement procedure also finds the correct answer for previous cases of overlapping peaks or isotope series in Figures 3 and 4, both by visual inspection and the criteria of largest adjusted R^2. Though the refinement procedure improves the results, it takes much longer to run. For the simple spectrum on the bottom left of Figure 4, refinement takes 20 s, whereas LASSO only takes less than 1 s on the same PC with Xeon CPU. Out of 10 spectrum segments described as examples, only 1 requires refinement to obtain the correct results. Therefore, for high-throughput analysis, the algorithm without refinement is preferred. If time is not a factor, we recommend the algorithm plus the refinement procedure.

Variable Selection as a Method for Spectrum Deconvolution. Variable selection is a natural choice for deconvolution of complex spectra. As a model selection technique for choosing competing models, variable selection aims at finding the smallest subset of variables to explain the data. This approach is not only intuitive but also is backed by vast statistical theories.[13] When one sees an isolated charge two isotope series, one simply interprets it as such, even though in theory an alternative explanation could be that it is a mixture of two charge one series 0.5 *m/z* apart from each other. Among the two competing explanations, one naturally chooses the simple one. In a sense, variable selection is already being used for manual interpretation of spectra. The LASSO method is efficient[15] and can handle cases where the number of variables is larger than the number of observations, i.e., the number of peaks. In contrast, an ordinary multiple regression method fails in such cases.

A common reason for multiple interpretations that are almost equally good is due to highly correlated variables. For spectrum deconvolution, it happens when the intensity-weighted sum of one or more isotope series looks like that of other isotope series. If two important variables are highly correlated, most variable

[15] Efron, B.; Johnstone, I.; Hastie, T.; Tibshirani, R. *Ann. Stat.* **2004**, *32* (2), 407−499.

selection methods, including LASSO, only select one of them. Sometimes the problem of correlated variable can be overcome by using quantized peptide mass distribution as a filter.

Using Quantized Peptide Mass Distribution as a Filter. The quantized mass distribution is obviously useful as a filter, though it is not necessary for examples in Figure 3. For example, the doubly charged series of 1000, 1000.5, 1001, and 1001.5 could be alternatively explained as the sum of two singly charged series of 1000, 1001 and 1000.5, 1001.5. However, according to the quantized peptide mass distribution, there are no peptide mass values within 0.2 Da of either 1000 or 999 Da. Therefore, this alternative explanation cannot be true. Today's instruments are capable of much better accuracy than 0.2 m/z. Obviously the quantized mass distribution would be more useful as a filter with increased mass accuracy.

It is a novel concept to use the quantized peptide mass distribution as a filter to aid spectrum deconvolution. Because the quantized distribution is due to the fractional parts of the mass of natural unmodified amino acids, which are between 0.01 and 0.1 Da, details of the virtual trypsin digestion such as the number of missed cleavages do not affect the fact that peptide mass distribution is quantized. Most chemical and biological modifications do not affect the quantized peptide mass distribution because they add masses that are within 0.1 Da of an integer and only affect certain specific amino acids. Such modifications include methonione oxidation, carbamidomethylated or acrylamide modified cysteines, phosphorylation, etc.

In three cases, the peptide mass distribution may not be quantized. First, in rare cases, the modification adds a noticeable mass defect that is greater than 0.2-Da deviation from the nearest integer. Second, the quantized mass distribution may not hold true if a significant fraction (e.g., >20%) of amino acids are modified by groups for which the fractional parts of the masses are greater than 0.8 Da, which are different from those of natural unmodified amino acids, i.e., 0.01−0.1 Da. Third, as the peptide mass increases, the quantized mass distribution gradually disappears. At mass values above 3000 Da, the mass distribution becomes almost continuous with bin sizes of ~0.1 Da. However, because most tryptic peptides are below 3000 Da, it is not a major problem. A sure way to check for quantized mass distribution is to perform virtual digestion on a sequence database and visually inspect the peptide mass distribution.

Using the quantized mass distribution as a filter is analogous to using the peptide mass distribution as the prior distribution of peptide masses in Bayesian statistics. Historically, the nonuniform distribution of peptide masses is already being used to calculate the MOWSE score[16] and by the widely used search engine MASCOT for protein identification.[17] However, for nonpeptide chemical species, the quantized peptide mass distribution should be used with caution.

Treatment of Nonpeptide Ions in the Spectrum. Ions that are considered chemical noise such as the solvent ions in ESI or matrix ions in MALDI have an atomic composition different from that of peptides; thus, their isotope patterns do not follow the averagine model calculated from peptides. Therefore, they could be a source of error, though the averagine model is still sometimes applicable. A good way to avoid such errors is to know the mass, isotope pattern, or charge states of such nonpeptide ions and identify them first before deconvolution is performed. This is possible; for example, in ESI-MS, many solvent ions are known to be singly charged and in the low-mass region.

CONCLUSIONS

We have proposed a new deconvolution algorithm for the mass spectra of complex peptide mixtures. Our algorithm is novel in three aspects. First, it formulates the deconvolution problem of complex spectra as a variable selection problem in the linear regression context, which is not only natural but also based on well-tested statistical theories. The linearity ensures the deconvolution is very efficient. Second, the algorithm makes use of the quantized peptide mass distribution to aid spectrum deconvolution, which complements the variable selection methods. Third, it explicitly accounts for errors in the expected isotope pattern (i.e., averagine) by using the importance value as a filter. The performance of the algorithm is tested with annotated spectra of peptide mixtures. The results are generally satisfactory except for one spectrum segment where an extra refinement step is necessary to obtain the correct results. Results are much better than those from BioAnalyst. Though the algorithm is only tested with ESI spectra, we expect it to be also useful for MALDI spectra, which are much simpler since peptides are singly charged. We hope this algorithm will be a useful tool in mass spectrometry-based proteomics research.

ACKNOWLEDGMENT

This work is supported by NIH-CA101150 (R.H.A.). We thank Dr. Jihyeon Lim for providing the serum spectra, and Dr. Sudha Rajagopalan for providing the ESI spectrum of 16 peptides. We thank Mr. Edward Nieves, Professor Charles B. Hall, and Dr. Christopher Paulse for reading the manuscript or providing helpful discussions.

SUPPORTING INFORMATION AVAILABLE

Appendix 1: derivation of equations for variable selection and choice of parameters. Appendix 2: definition of adjusted R^2. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review December 14, 2005. Accepted March 10, 2006.

AC052212Q

[16] Pappin, D. J. C.; Hojrup, P.; Bleasby, A. J. *Curr. Biol.* **1993**, *3* (6), 327−332

[17] Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20* (18), 3551−3567