

Shotgun Protein Sequencing by Tandem Mass Spectra Assembly

Nuno Bandeira,* Haixu Tang, Vineet Bafna, and Pavel Pevzner

Computer Science and Engineering Department, University of California, San Diego, Department 0114, 9500 Gilman Drive, La Jolla, California 92093-0114

The analysis of mass spectrometry data is still largely based on identification of *single* MS/MS spectra and does not attempt to make use of the extra information available in *multiple* MS/MS spectra from partially or completely overlapping peptides. Analysis of MS/MS spectra from multiple overlapping peptides opens up the possibility of assembling MS/MS spectra into entire proteins, similarly to the assembly of overlapping DNA reads into entire genomes. In this paper, we present for the first time a way to detect, score, and interpret overlaps between uninterpreted MS/MS spectra in an attempt to sequence entire proteins rather than individual peptides. We show that this approach not only extends the length of reconstructed amino acid sequences but also dramatically improves the quality of de novo peptide sequencing, even for low mass accuracy MS/MS data.

Traditional MS/MS-based protein analysis starts from a specific digestion of a protein into *nonoverlapping* (usually tryptic) peptides. The nonspecific digestion into *overlapping* peptides is hardly ever used in MS/MS studies, and the common perception is that nonspecific digestion only complicates the already difficult protein identification problem and should be avoided. However, in a pioneering experiment back in 1989 Hopper et al.¹ took advantage of spectra from overlapping peptides to de novo sequence a whole protein from the rabbit bone marrow. Today, it is feasible to run experiments where the proteins are separately digested with different enzymes such as trypsin and pepsin, resulting in the acquisition of MS/MS data from more partially or completely overlapping (i.e., identical) peptides from the proteins in the sample. These types of data have a clear parallel with the type of data obtained in whole genome sequencing where overlapping DNA reads were collected and assembled into whole genomes. However, it is not clear how to take advantage of overlapping spectra in MS/MS analysis, and in the 15 years since the Hopper et al. paper,¹ there was no attempt to assemble uninterpreted spectra from overlapping peptides. In this paper, we show that MS/MS spectra assembly is feasible and demonstrate that it leads to a highly accurate approach to de novo sequencing of entire proteins.

The feasibility of generating and the benefits of using rich peptide ladders were recently demonstrated in two different

contexts. Woods and co-workers^{2–9} demonstrated that rich peptide ladders can be generated by nonspecific proteolytic digestion in the context of hydrogen exchange (DXMS) studies of protein structure. Yates and co-workers¹⁰ recognized the potential of nonspecific proteolytic digestion in improving the procedures for database search of posttranslationally modified proteins. In the latter, the richer set of peptides generates enough MS/MS spectra from nonmodified peptides to create a smaller protein sequence database that is then searched for posttranslational modifications. Promising results were presented, but the methodology faces difficulties in that it depends on having at least one (or more for reliable identification) good MS/MS spectrum from an unmodified peptide to first identify the protein in the sample. Moreover, there is a delicate balance between choosing too many protein candidates or choosing less candidates but taking the risk of not including the correct protein sequence in the subsequent search for posttranslational modifications. Neither of these approaches attempts to assemble noninterpreted MS/MS spectra.

In this pilot experiment, we capitalize on the principle “*pairwise alignments whisper while multiple alignments shout out loud*” that was well explored in genomics but so far has not been applied to MS/MS studies. Our approach provides a proof of concept by showing that the de novo interpretation of unknown protein sequences can be significantly improved by detecting overlaps between uninterpreted MS/MS spectra to increase the quality and extent of de novo interpretations. By making absolutely *no* use of any database information, we avoid the pitfalls of current methods in that we do not require knowledge of the protein sequence and do not face the same exponential growth in running time when

- (2) Pantazatos, D.; Kim, J. S.; Klock, H. E.; Stevens, R. C.; Wilson, I. A.; Lesley, S. A.; Woods, V. L. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (3), 751–756.
- (3) Hamuro, Y.; Anand, G. S.; Kim, J. S.; Juliano, C.; Stranz, D. D.; Taylor, S. S.; Woods, V. L. *J. Mol. Biol.* **2004**, *340* (5), 1185–1196.
- (4) Black, B. E.; Foltz, D. R.; Chakravarthy, S.; Luger, K.; Woods, V. L.; Cleveland, D. W. *Nature* **2004**, *430* (6999), 578–582.
- (5) Englander, J. J.; Del Mar, C.; Li, W.; Englander, S. W.; Kim, J. S.; Stranz, D. D.; Hamuro, Y.; Woods, V. L. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (12), 7057–7062.
- (6) Woods, V. L.; Hamuro, Y. *J. Cell. Biochem.* **2001**, (Suppl 37), 89–98.
- (7) Pantazatos, D.; Kim, J. S.; Klock, H. E.; Stevens, R. C.; Wilson, I. A.; Lesley, S. A.; Woods, V. L., Jr. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (3), 751–6.
- (8) Hamuro, Y.; Zawadzki, K. M.; Kim, J. S.; Stranz, D. D.; Taylor, S. S.; Woods, V. L. *J. Mol. Biol.* **2003**, *327* (5), 1065–1076.
- (9) Hamuro, Y.; Burns, L.; Canaves, J.; Homan, R.; Taylor, S.; Woods, V. J. *Mol. Biol.* **2002**, *321* (4), 703–714.
- (10) MacCoss, M. J.; McDonald, W. H.; Saraf, A.; Sadygov, R.; Clark, J. M.; Tasto, J. J.; Gould, K. L.; Wolters, D.; Washburn, M.; Weiss, A.; Clark, J. I.; Yates, J. R. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (12), 7900–7905.

* To whom correspondence should be addressed. E-mail: bandeira@cs.ucsd.edu.

(1) Hopper, S.; Johnson, R. S.; Vath, J. E.; Biemann, K. *J. Biol. Chem.* **1989**, *264* (34), 20438–20447.

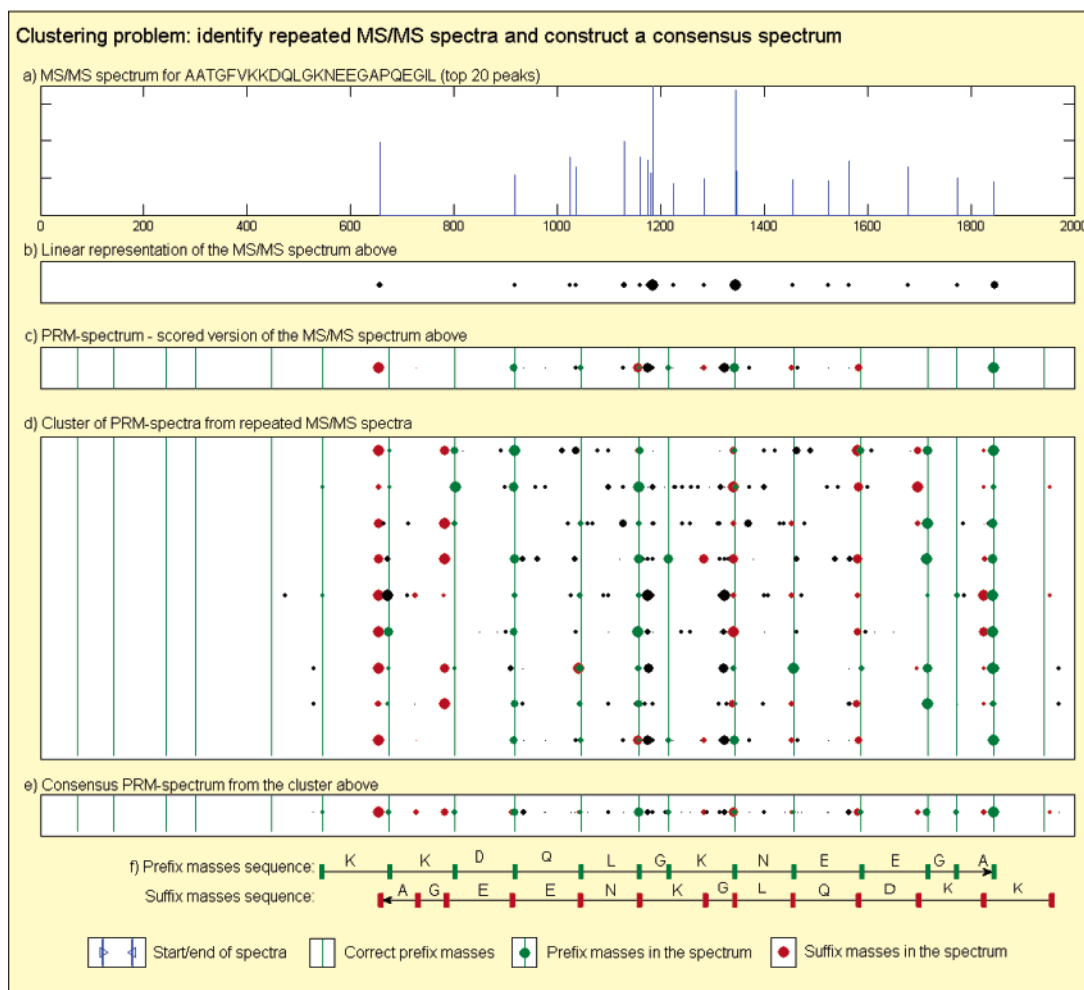


Figure 1. Clustering phase. (a) and (b) illustrate our linear representation of spectra where a dot indicates a peak and the dot size is proportional to the peak height (used to save space when showing multiple alignments of several spectra). (c) shows the corresponding PRM spectrum (our preprocessed and scored version of an MS/MS spectrum). For the convenience of the reader, prefix masses are shown in green, and suffix masses are shown in red, although this distinction is not known in advance. Other masses (which do not correspond to prefix or suffix masses) are shown as black dots. (d) Clustering is then used to take advantage of redundant information in multiple spectra from the same peptide and (e) obtain a single, more reliable, consensus PRM spectrum (some of the red dots are hidden by green dots). All black dots still present in (e) correspond either to neutral losses or to doubly charged fragments. The increased number and significance of red/green dots in the consensus PRM spectrum as compared to individual spectra would already yield a reliable de novo peptide sequence (as illustrated in (f)), although we refrain from interpreting the spectra until the end of the assembly phase (Figure 3).

considering posttranslational modifications. Experimental results are provided using a data set of 2646 α -synuclein MS/MS spectra, 303 of which were identified as 83 overlapping α -synuclein peptides. Hopefully our proof of concept will further show the potential of using nonspecific digestion enzymes in proteomics experiments and promote the availability of more and larger such data sets.

Similarly to the overlap \rightarrow layout \rightarrow consensus approach in DNA fragment assembly, we propose a clustering \rightarrow alignment \rightarrow layout \rightarrow de novo interpretation approach for MS/MS analysis (Figures 1–3). The first *clustering* stage includes the generation of prefix residue mass (PRM) spectra (scored version of the MS/MS spectra) and the clustering procedure to detect repeated spectra from the same peptide and build the corresponding consensus spectra (Figure 1). In the second *alignment* stage (Figure 2), we address the pairwise alignment of PRM spectra to detect overlaps and describe the construction of the overlap graphs. Our third *assembly* stage uses the overlap graph to

assemble spectra and subsequently determine the best amino acid sequence (Figure 3).

PREFIX RESIDUE MASS SPECTRA

A peptide can be defined as a string $\rho = a_1, \dots, a_n$, where a_i is any amino acid with a known residue mass $m(a_i)$. Also, any prefix $\rho_i = a_1, \dots, a_i$ has a PRM $m(\rho_i) = \sum_{j=1}^i m(a_j)$; the special case $m(\rho) = m(\rho_n)$ is also referred to as *parent mass*. As such, an equivalent representation of a peptide $\rho = a_1, \dots, a_n$ is given by the mass series

$$\mathcal{R} = \{m(\rho_1), \dots, m(\rho_n)\}$$

Another equivalent representation is given by the reverse mass series \mathcal{R}^{REV} —the masses of all suffixes of ρ

$$\mathcal{R}^{\text{REV}} = \{m(\rho) - m(\rho_{n-1}), \dots, m(\rho) - m(\rho_1), m(\rho)\}$$

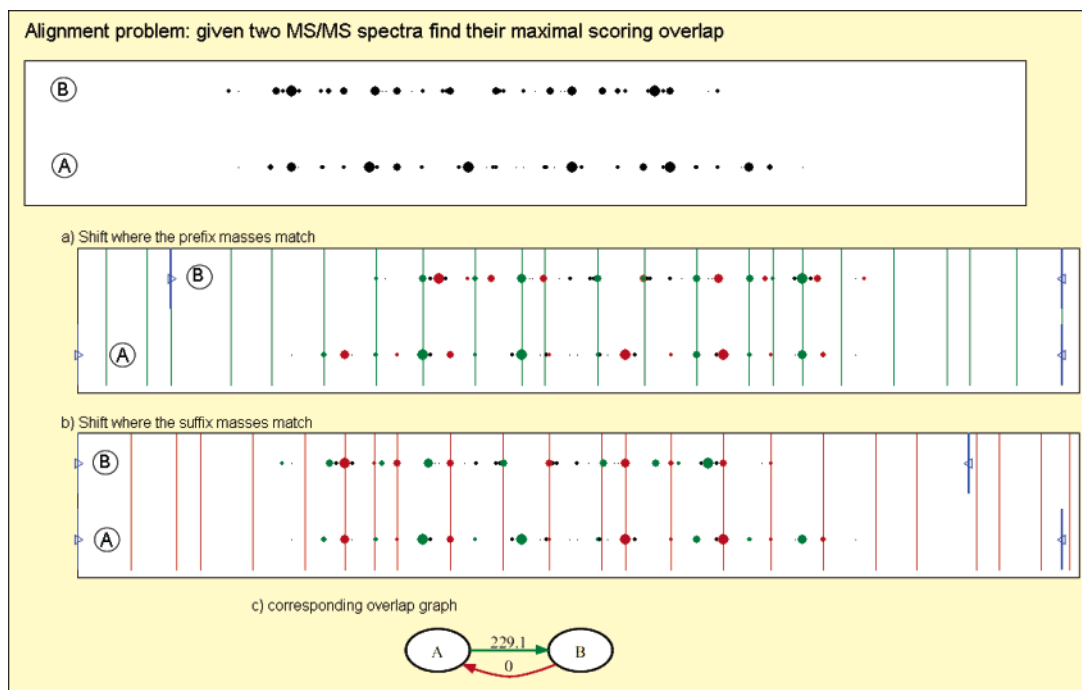


Figure 2. Pairwise alignment phase. The two MS/MS spectra used in this example correspond to the peptides ATGFVKKDKLGNKEEGAPQEGIL (spectrum A) and FVKKDKLGNKEEGAPQEGIL (spectrum B); the peptides themselves are unknown to the algorithm. This example illustrates the case where one peptide is contained in the other, but it does not have to be so; our method detects partially overlapping spectra as well. When aligning two PRM spectra, we look for a maximal scoring shift between them (score is proportional to the number of matched PRMs). Since PRM spectra are symmetric, we always have two such shifts: (a) the shift 229.1 (total mass of ATG) where the prefix masses match; (b) the shift 0 where the suffix masses match (peptide B is a suffix of peptide A). The resulting pairwise alignment is represented as two edges between vertexes A and B as shown in (c) where the shift with matching prefix (suffix) masses is shown in green (red). This representation is further used in the assembly phase (Figure 3).

In general, given a set of masses $X = \{x_1, \dots, x_m\}$ with associated parent mass $m(X)$, we define the reverse of X as $X^{\text{REV}} = \{m(X) - x_m, \dots, m(X) - x_1, m(X)\}$ and the λ shift of X as $X^\lambda = \{x_1 + \lambda, \dots, x_m + \lambda\}$.

Using this setup, a theoretical MS/MS spectrum S for a peptide ρ is defined as

$$S = \mathcal{R}^{-1} \cup (\mathcal{R}^{\text{REV}})^{\lambda}$$

where elements of \mathcal{R}^{-1} and $(\mathcal{R}^{\text{REV}})^{\lambda}$ correspond to b and y ions, respectively [b ions include an additional H atom (+1 Da); y ions include an H₂O molecule (+18 Da) from the C-terminal and also an additional H atom (+1 Da) for a total peak offset of +19 Da.] (neutral losses are considered later while scoring the spectrum). Also, we denote the parent mass of a spectrum S from a peptide ρ as $m(S) = m(\rho)$.

In mass spectrometry, one often faces the inverse problem of transforming an experimental spectrum S into the mass series representation of a peptide. The simplest approach to this inverse problem is to reverse the transformation above

$$\text{PRM}(S) = S^{-1} \cup (S^{-19})^{\text{REV}}$$

The set $\text{PRM}(S)$ represents an attempt to reconstruct the set $\mathcal{R} \cup \mathcal{R}^{\text{REV}}$ of the peptide ρ that generated S and defines the peak positions (PRMs) in our PRM spectrum. Figure 4 illustrates the steps described above. Ideal PRM spectra could be built from MS/MS spectra containing only b and y ion peaks. This ideal setup

can be approximated by selecting peaks from the experimental MS/MS spectra according to their intensity—higher intensity peaks tend to correspond to b and y ion peaks. As such, PRM positions in a PRM spectrum were determined using only the top 20 highest intensity peaks in the MS/MS spectrum. The choice to keep 20 peaks per MS/MS spectrum was motivated by the analysis of peak annotation histograms, which show a very low percentage of b/y ion peaks outside the top 20 intensity peaks (data not shown); b/y ion peaks are the most important peaks in determining the correct positions for the PRMs, which are then scored using all the peaks in the MS/MS spectrum.

Every peak s in an MS/MS spectrum S generates two PRMs ($s - 1$ and $m(S) - s + 19$) called *twin* PRMs. Every PRM spectrum P is then necessarily symmetric because every pair of twin PRMs is symmetric about $(M(P)/2)$, where $M(P)$ abbreviates $m(P) + 18$.

Scoring PRMs in PRM spectra Not all PRMs are created equal—some have more compelling evidence of being correct than others by having, for example, both corresponding b and y ions and neutral loss peaks present in the MS/MS spectrum. To reflect how confident we are in a PRM, we use the Dančik et al.¹¹ scoring scheme. (Readers familiar with the scoring defined in that paper may recognize the connection between PRM spectra and scored vertexes in the spectrum graph.) The details of our application of this scoring scheme can be found in Supporting Information (see refs 12–15 for other applications of the same scoring scheme).

(11) Dančik, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. *J. Comput. Biol.* **1999**, *6* (3–4), 327–342.

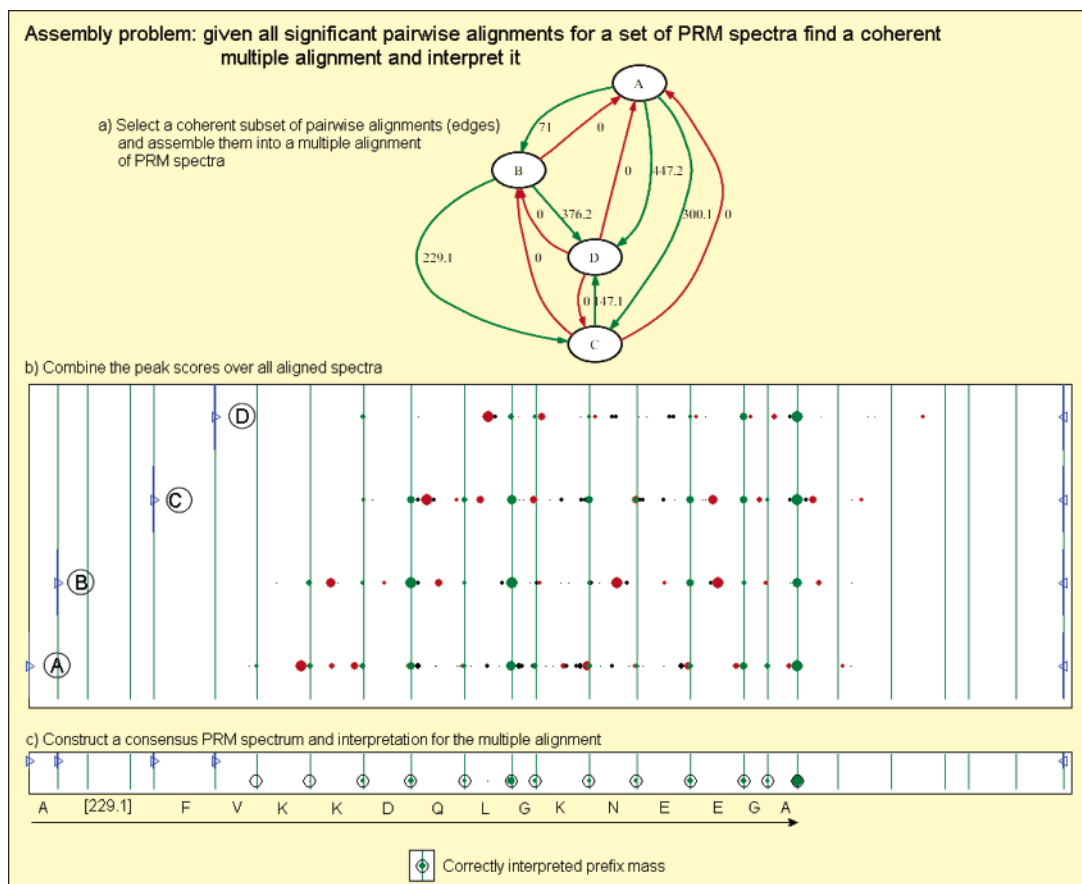


Figure 3. Assembly phase. (a) shows the overlap graph constructed for four PRM spectra where edges represent optimal pairwise alignments. Our assembly algorithm finds the optimal coherent subset of edges that defines the path $A \xrightarrow{71} B \xrightarrow{229.1} C \xrightarrow{147.1} D$. The edges $A \xrightarrow{300.1} C$, $A \xrightarrow{447.2} D$, and $B \xrightarrow{376.2} D$ provide additional support for this path ($300.1 = 71 + 229.1$, $376.2 = 229.1 + 147.1$, $447.2 = 71 + 376.2$) and are thus also included in the selected set of edges (the green edges). The corresponding multiple alignment shown in (b) is used to construct the consensus PRM spectrum shown in (c) and recover the indicated amino acid sequence. De novo interpretation of the assembled MS/MS spectra becomes much simpler because noise was completely removed from the consensus PRM spectrum.

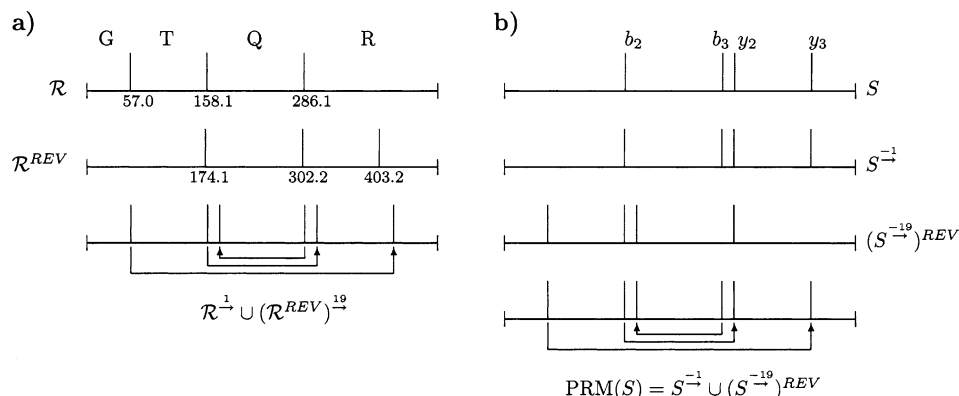


Figure 4. Obtaining the PRM spectrum. Part a illustrates how the sets \mathcal{R} and \mathcal{R}^{REV} are used to define the theoretical PRM spectrum for the peptide GTQR. Part b shows how a hypothetical MS/MS spectrum S for GTQR is processed to obtain the PRM spectrum $PRM(S)$. In both (a) and (b), arrows indicate the prefix/suffix pairs.

This is particularly adequate for our purposes because it allows us to score putative PRMs without having a putative peptide interpretation (as is the case in database search). Some approaches also take into consideration the relative intensity values

of the observed ion types^{16–19} although the best way to incorporate such information is still an open problem under consideration.^{20–22}

The scores defined by Dančik et al. are additive due to a log scaling and have the expected positive premium and negative penalty score changes for ion types with probability of occurrence higher than the probability of background noise. Each PRM $p_i \in P$ is thus assigned a weight $w(p_i)$ by looking for supporting ion peaks in the corresponding MS/MS spectrum S to obtain a PRM

(12) Lubeck, O.; Sewell, C.; Gu, S.; Chen, X.; Cai, D. *Proc. IEEE* **2002**, *90* (12), 1868–1874.

(13) Havilio, M.; Haddad, Y.; Smilansky, Z. *Anal. Chem.* **2003**, *75* (3), 435–444.

spectrum $P = \{p_1, \dots, p_n\}$ having associated weights $\{w(p_1), \dots, w(p_n)\}$. In the following, we assume that in every PRM spectrum the PRMs are sorted by increasing mass.

CLUSTERING MS/MS SPECTRA

Repeated MS/MS spectra (multiple spectra from the same peptide) are common in high-throughput MS/MS experiments. Recent approaches either attempt to discard these to speed up database searches²³ or average over the multiple copies to increase the intensity of correct peaks relative to noise peaks²⁴ (although averaging could retain high intensity noise peaks in the consensus spectrum). A recent study by Venable and Yates²⁵ on the variability of experimental MS/MS spectra from the same peptide provides evidence that peak intensities vary considerably between repeated MS/MS spectra and also argues that although MS/MS spectra averaging improves database search results other approaches may perform better.

We propose to use the redundant information in the repeated MS/MS spectra to filter noise based solely on the principle that real MS/MS spectrum peaks should be present in most MS/MS spectra from the same peptide and the randomly distributed noise peaks should not. But to make use of the redundant information in independently obtained PRM spectra of the same peptide, we first need to decide whether two PRM spectra originate from the same peptide using *only* the information contained in the PRM spectra.

A naive approach to this problem could be based on the shared peak count between two spectra. However, this ignores the fact that some peaks in an MS/MS spectrum have more evidence of being true peaks than others, for example, by having additional peaks at corresponding neutral loss positions. It may also happen that in one MS/MS spectrum we only observe a b ion for a given fragmentation point and in the other MS/MS spectrum we only observe a y ion for the same fragmentation point, in which case there are no matching peaks although there is relevant matching information in the spectra. Matching PRM spectra instead of MS/MS spectra addresses both of these points. A match between two PRM spectra P and Q can then be defined as a set $P \cap Q$ of matching PRMs (p_i, q_j) with associated weights $w(p_i) + w(q_j)$. The weight of any set of PRMs $X = \{x_1, \dots, x_n\}$ is simply given by $w(X) = \sum_{x_i \in X} w(x_i)$.

Matching PRM Spectra: Sparse Subsets. A subset of a PRM spectrum P is called *sparse* if no two PRMs are less than 57

Da apart (i.e., the mass of the lightest amino acid glycine). In PRM spectra, peaks are supposed to correspond to prefix residue masses. Therefore, closely located PRMs (i.e., less than 57 Da apart) cannot both be correct. These closely positioned PRMs can be avoided by finding sparse subsets of PRMs.

To find a maximum weight sparse subset of a PRM spectrum $P = \{p_1, \dots, p_n\}$, we define a simple dynamic programming recursion where $D(i)$ is the maximum weight of a sparse subset of $\{p_1, \dots, p_i\}$ that includes p_i . Then

$$D(i) = w(p_i) + \max_{j:p_j \leq p_i - 57} D(j)$$

Matching PRM Spectra: Antisymmetric Subsets. Although computing maximum weight sparse subsets would already impose tighter conditions for PRM spectra matching, it may still happen that MS/MS spectrum peaks are double counted in the matching process. As described in the previous section, an MS/MS spectrum peak $s \in S$ generates a pair of twin PRMs: $s - 1$ and $m(S) - s + 19$. Since both these PRMs are scored using the same MS/MS spectrum peaks, including both PRMs in a match effectively counts the same MS/MS spectrum peaks twice and should be avoided. As such, a subset X of a PRM spectrum P is defined as *antisymmetric* if it has no twin PRMs, i.e., no two PRMs in X add up to $M(P)$.

Matching PRM Spectra: Optimal Subsets. A subset of a PRM spectrum P is *optimal* if it is a sparse and antisymmetric subset of maximum weight. Computing an optimal subset of a set of PRMs is algorithmically the same problem as the de novo problem of finding the peptide that best explains a spectrum.^{11,26,27} The only difference between these two problems is that in the latter there is only a limited set of valid jumps between PRMs (corresponding to amino acid masses) and in the former any jump of ≥ 57 Da is a valid jump. A detailed description of the implemented algorithm for the computation of optimal subsets can be found in section A.2 of the Supporting Information.

Match Score between PRM Spectra. An optimal match between two PRM spectra P and Q is simply an optimal subset of their overlap $P \cap Q$. Although the weight of an optimal match between PRM spectra is already a good measure of similarity, we observed that sometimes spurious high-scoring matches occur when only a few PRMs match in a small mass range, simply by chance or due to local sequence similarities. On the other hand, repeated PRM spectra from the same peptide tend to match most high-scoring PRMs in a large mass range. To account for this effect, we introduce a correction factor α —the percentage of mass range covered by the restricted match. Using d_{PQ} as the difference between the maximum and minimum masses of the matched PRMs (i.e., match range) and m_{PQ} as the parent mass of the matched PRM spectra, this correction factor is thus defined as $\alpha = (d_{PQ}/m_{PQ})$.

The *match score* \mathcal{M} between P and Q is then defined as $\mathcal{M} = \alpha \times w(Y)$, where Y is an optimal match between P and Q .

Constructing Clusters of PRM Spectra. After computing the match scores \mathcal{M} , we consider two spectra as similar if their match

(14) Colinge, J.; Magnin, J.; Dessingy, T.; Giron, M.; Masselot, A. *Proteomics* **2003**, *3* (8), 1434–1440.

(15) Cannon, W. R.; Jarman, K. D. *Rapid Commun. Mass Spectrom.* **2003**, *17* (15), 1793–1801.

(16) Eng, J. K.; McCormack, A. L.; Yates, J. R. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (11), 976–989.

(17) Yates, J. R.; Eng, J. K.; McCormack, A. L. *Anal. Chem.* **1995**, *67* (18), 3202–3210.

(18) Tabb, D. L.; Saraf, A.; Yates, J. R. *Anal. Chem.* **2003**, *75* (23), 6415–6421.

(19) Elias, J. E.; Gibbons, F. D.; King, O. D.; Roth, F. P.; Gygi, S. P. *Nat. Biotechnol.* **2004**, *22* (2), 214–219.

(20) Schutz, F.; Kapp, E. A.; Simpson, R. J.; Speed, T. P. *Biochem. Soc. Trans.* **2003**, 1479–1483.

(21) Reference deleted in proof.

(22) Tabb, D. L.; Smith, L. L.; Brechi, L. A.; Wysocki, V. H.; Lin, D.; Yates, J. R. *Anal. Chem.* **2003**, *75* (5), 1155–1163.

(23) Tabb, D. L.; MacCoss, M. J.; Wu, C. C.; Anderson, S. D.; Yates, J. R., 3rd. *Anal. Chem.* **2003**, *75* (10), 2470–7.

(24) Beer, I.; Barnea, E.; Ziv, T.; Admon, A. *Proteomics* **2004**, *4* (4), 950–960.

(25) Venable, J. D.; Yates, J. R. *Anal. Chem.* **2004**, *76* (10), 2928–2937.

(26) Chen, T.; Kao, M. Y.; Tepel, M.; Rush, J.; Church, G. M. *J. Comput. Biol.* **2001**, *8* (3), 325–337.

(27) Bafna, V.; Edwards, N. *Proc. 7th Annu. Int. Conf. Comput. Mol. Biol.* 2003; pp 9–18.

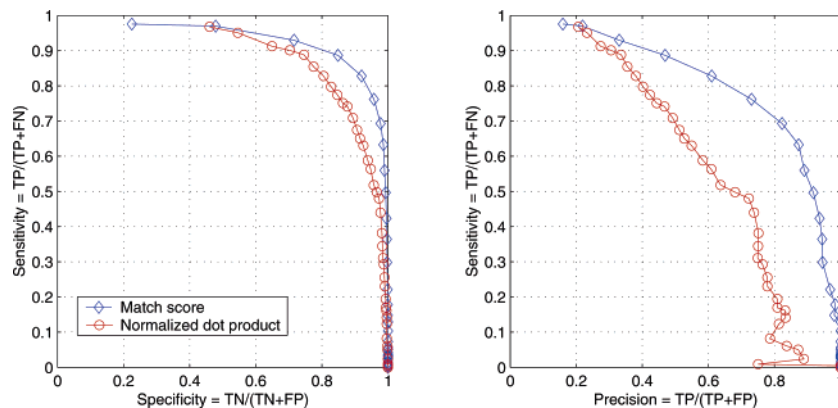


Figure 5. ROC curve (left) and precision vs sensitivity (right). The triangle condition is enforced on both methods.

score is above the chosen threshold. A natural extension of the pairwise similarity concept when considering clusters of PRM spectra is to require each spectrum to coherently match at least two other spectra, which must also match each other. A simple example of a cluster rejected through this condition is a starlike cluster of size n , where $n - 1$ spectra match only a single spectrum. This requirement is thus enforced as the *triangle* condition: a match between PRM spectra P and Q is retained if and only if there is some other PRM spectrum R such that P matches R and Q matches R . Clearly this step removes any cluster of PRM spectra of size <3 . The remaining matches define connected components interpreted as clusters.

Clustering Results. To evaluate the performance of our clustering procedure, we used a set of 1455 Sequest annotated MS/MS spectra (the set of experimental MS/MS spectra is described under Experimental Results). Sequest annotations were used only for validation purposes and are *not* used by our algorithm at any point. A match between two PRM spectra is considered correct if the peptide annotations are the same for the MS/MS spectra originating the matched PRM spectra. Every spectrum was matched against every other spectrum with a parent mass difference not exceeding 2 Da; on average each spectrum was matched against 7.3 other spectra.

Figure 5 shows how true/false positives (TP/FP) and true/false negatives (TN/FN) vary for different thresholds on the match score \mathcal{M} ; a receiver operating characteristic (ROC) curve is shown on the left, and because the number of true positives is only $\sim 10\%$ of the number of true negatives, the precision versus sensitivity curve is also shown on the right. For comparison purposes, Figure 5 also includes curves for the normalized dot-product approach proposed in^{23,24} as a similarity metric between MS/MS spectra.

As shown in Figure 5, our method clearly outperforms the normalized dot-product approach. One possible reason our match score approach performs better than the normalized dot-product is the variability in peak intensity between different MS/MS spectra of the same peptide (recently studied by Venable and Yates²⁵). The match score \mathcal{M} thus allows us to separate between correct and incorrect pairwise matches with the choice of adequate threshold conditioned by the instrument parameters and the level of different peptides with the same precursor mass expected from the experiment. For our alignment and assembly purposes, we selected a subset of matches as detailed in Table 1.

Sequest peptide annotations were also used to estimate the quality of the clusters obtained—the median percentage of non-

Table 1. Match Results in the Clustering Phase

	no. of matches	no. of correct matches	% correct
total	5322	697	13%
after thresholding M	823	545	66%
after triangle condition	617	501	80%

matching peptide annotations in a cluster was found to be 11%. The retained 617 matches result in a total of 39 clusters—29 annotated as coming from the protein in our sample and 10 where the peptide annotations do not match. While it is possible that these 10 clusters are retained because our match score threshold was not aggressively selective, it may also be the case that the annotations are incorrect—only the highest scoring peptide annotation was retained from the database search procedure. In any event, the clustering procedure can be made as selective as desired (depending on the experiment requirements), with an acceptable penalty in sensitivity. Our choice of sensitivity/selectivity tradeoff reflects the fact that the obtained clusters are not our final goal but rather a preprocessing step for the alignment procedure where some amount of noise (incorrect PRM spectra) is tolerable. Also, a minor amount of incorrect MS/MS spectra in any single cluster does not produce a significant amount of noise in the corresponding consensus spectrum (see next section).

Building Consensus PRM Spectra. The usefulness of any spectral clustering technique is defined by how well the consensus spectrum reflects the true peaks in all spectra originating from the same peptide. As mentioned above, our approach to this problem is to score the putative PRMs across all clustered spectra—real peaks should appear in most MS/MS spectra (albeit with varying intensities) and noise peaks should not. As such, when given a cluster $C = \{P_1, \dots, P_k\}$, a single consensus PRM spectrum can be constructed by a direct extension of the scoring procedure defined above. The weight $w(t, C)$ of a putative PRM t over the cluster C is given by

$$w(t, C) = \sum_{i=1}^k w(t, P_i)$$

where $w(t, P_i)$ is the PRM weight (positive or negative) for the mass t in the i th PRM spectrum in the cluster. Negative PRM scores occur whenever there is little or no evidence that a putative

Table 2. Quality of the PRMs in the Consensus PRM Spectra

type of fragment originating PRM	median % of PRM spectrum score
b/y	77
neutral loss or doubly charged	10
unexplained	12

PRM represents a real prefix residue mass, for example, by not having corresponding b or y ion peaks in the MS/MS spectrum. This is a common event when scoring a putative PRM t originating in a noise peak in one of the clustered spectra—most other spectra will have no peaks supporting t and the overall score for t will thus be negative. The consensus PRM spectrum for a cluster considers all putative PRMs in all PRM spectra in the cluster but retains only the PRMs with a positive summed score. In this way, PRMs generated by high-intensity unexplained peaks in any MS/MS spectrum are not likely to be present in the consensus PRM spectrum because its absence in all other spectra will make its summed score negative. Although relative intensities vary across multiple MS/MS spectra from the same peptide²⁵ the presence or absence of real fragment peaks tends to be stable. As shown in Table 2 our resulting consensus PRM spectra are dominated by high-scoring PRMs at the correct prefix and suffix mass positions, and almost half of the remaining PRMs correspond to either doubly charged fragment masses or neutral losses (Figure 1). As a result, the de novo interpretation of the consensus PRM spectra is greatly simplified as compared to individual spectra. However, we refrain from de novo interpretation at this stage to take advantage of the assembled MS/MS spectra that we describe below.

ALIGNING MS/MS SPECTRA

The purpose of PRM spectrum alignment is to determine how much overlap, if any, exists between two peptides given only two uninterpreted PRM spectra, one from each peptide. When a large overlap exists, then there is some shift of the PRMs in one PRM spectrum such that these match the PRMs in the other PRM spectrum and the sum of the scores of the matched PRMs is high (Figure 2).

Every shift λ between two PRM spectra P and Q defines a partial overlap region with a corresponding set of matching PRMs ($P \cap Q^{\lambda}$). As such, scoring a shift is almost the same as scoring full spectrum matches as covered in the clustering section. The only difference is that the requirement to exclude twin PRMs can now be dropped because these are not expected to match simultaneously in partial overlaps. Thus, in this context, it suffices to compute a maximum weight sparse subset Y of ($P \cap Q^{\lambda}$) and set the shift score to $w(Y)$. Moreover, due to the inherent symmetry of MS/MS and PRM spectra, every shift λ has a *symmetric* shift λ_S with exactly the same score; in correct alignments one of the shifts matches the prefix masses and its symmetric shift matches the suffix masses (Figure 2). The center of symmetry when aligning P to Q is given by $c = (m(P) - m(Q))/2$, and as such, any shift λ has a symmetric shift λ_S given by $\lambda_S = 2c - \lambda$. Therefore, the best alignment between two PRM spectra is defined not by a single shift but by a pair of symmetric shifts (λ , λ_S).

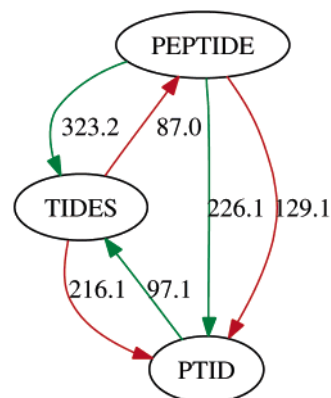


Figure 6. Example overlap graph. Each vertex represents a PRM spectrum from the listed peptide and edges represent shifts corresponding to the highest scoring alignment (red/green pairs) between spectra. For example, 323.2 corresponds to the mass of PEP while 87.0 corresponds to the mass of S.

Overlap Graph. The best alignments between PRM spectra define a directed *overlap* graph where each vertex corresponds to a PRM spectrum and each edge corresponds to a shift between two PRM spectra. Only the highest scoring alignment is used to define edges between two vertices and edge directionality is used to represent the sign of the shifts: a positive shift λ from P to Q defines an edge (P, Q) , and a negative shift λ' from P to Q defines an edge (Q, P) . Every edge $e = (P, Q)$ is characterized by $\lambda(e)$ (the shift between P and Q) and $w(e)$ (the shift score as defined above). Additionally, green edges represent the shifts when prefix masses match and red edges represent the symmetric shifts when suffix masses match. Figure 6 shows an example of an overlap graph on three imaginary PRM spectra from the peptides listed in the vertices.

Filtering Edges in the Overlap Graph. Since we compute alignments for all pairs of PRM spectra and every such pair will have some best symmetric shift pair, we are bound to have many incorrect pairwise alignments that need to be filtered. We address this issue by building on the principle that a correct alignment should match most of the high scoring PRMs in the overlap region and define a quality score β as the ratio between the matched and unmatched PRM scores.

Given a pair of PRM spectra P and Q for which the best alignment is (λ, λ_S) , let $M_\lambda (M_{\lambda_S})$ be the maximum weight sparse subset of $P \cap Q^{\lambda}$ ($P \cap Q^{\lambda_S}$) and let $M = M_\lambda \cup M_{\lambda_S}$. Conversely, let U_P be the set of all the unmatched PRMs in the overlapped regions of P when shifting by λ and λ_S . The quality score is then defined as $\beta_P = (w(M)/w(U_P))$ (similarly for β_Q). Figure 7 shows the ROC and precision/sensitivity curves obtained by varying a threshold t and selecting edges $e = (P, Q)$ from the overlap graph where both $\beta_P \geq t$ and $\beta_Q \geq t$.

As in the clustering section, the triangle condition is also enforced in the overlap graph but only whenever applicable. A *valid triangle* is defined by three edges $e_{PQ} = (P, Q)$, $e_{QR} = (Q, R)$, and $e_{PR} = (P, R)$ if $\lambda(e_{PR}) = \lambda(e_{PQ}) + \lambda(e_{QR})$ and is invalid otherwise. Therefore, if an edge in the overlap graph is part of some triangle, then either it belongs to at least one valid triangle or it is removed from the overlap graph. If the edge does not belong in some triangle (e.g., a set of only two PRM spectra) then this restriction does not apply.

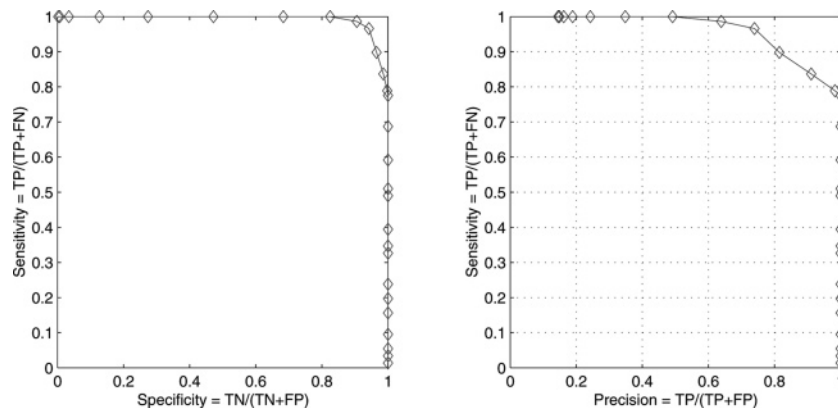


Figure 7. ROC curve (left) and precision vs sensitivity (right) for selection of pairwise alignments.

Table 3. Pairwise Alignment Results; Two Symmetric Shifts Per Pairwise Alignment

	no. of shifts	no. of correct shifts	% correct
total	$2 \times 39 \times 38 = 2964$	147	5.0
after filtering	114	114	100

Pairwise alignments were computed for all 39 PRM spectra obtained from the clustering phase as described in the previous section; results are shown in Table 3.

The 114 pairwise alignments define 5 connected components in the overlap graph and are the input to the assembly stage of our method.

ASSEMBLING MS/MS SPECTRA

The first step in going from an overlap graph to an assembly and interpretation of the partially overlapping spectra is to make the distinction between red and green edges as illustrated in Figure 6. In reality, after building the overlap graph, the colors of the edges are unknown.

Decomposing the Overlap Graph. In any acceptable solution to the assembly problem, each vertex in the overlap graph has a unique position in the assembled sequence. The conventional fragment assembly problem assigns a coordinate (e.g., a starting position in the genome) to every fragment while trying to optimize some target function. Similarly, the MS/MS assembly problem attempts to assign a coordinate to every MS/MS spectrum. The difference is that the coordinate of an MS/MS spectrum from a peptide starting at position i of a protein ρ_1, \dots, ρ_n corresponds to the mass of the first $i - 1$ amino acids. Thus, let G be an overlap graph and let Ψ be a function (called *vertex potential*) that assigns a coordinate to every vertex in G . Figure 8 shows an overlap graph with assigned vertex potentials.

An edge $e = (v, u)$ is called *coherent* with respect to a potential Ψ if $\lambda(e) = \Psi(u) - \Psi(v)$. The vertex potentials Ψ in Figure 8 define six coherent edges of overall weight $w(\Psi)$ given by

$$\begin{aligned}
 w(\Psi) &= w(A \rightarrow B) + w(A \rightarrow C) + w(A \rightarrow D) + \\
 &\quad w(B \rightarrow C) + w(B \rightarrow D) + w(C \rightarrow D) \\
 &= 3.0 + 3.5 + 2.9 + 2.4 + 1.8 + 2.2
 \end{aligned}$$

Given an overlap graph, we are interested in finding a potential Ψ of maximal weight $w(\Psi)$ —*maximal coherent edge-set problem*.

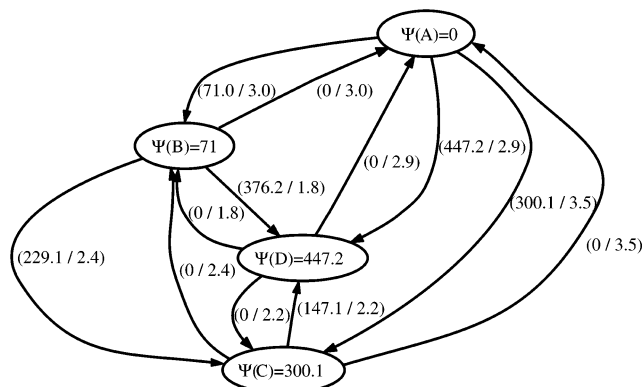


Figure 8. Overlap graph from Figure 2 with assigned vertex potentials $\Psi(v)$. Edges are labeled by (shift λ , shift score $w(\lambda)$) pairs.

Similarly to the DNA fragment assembly problem, the maximal coherent edge-set problem is NP-complete. However, the overlap graphs arising in MS/MS assembly are rather small (in contrast to overlap graphs arising in DNA fragment assembly) rendering the MS/MS assembly problem simpler in practice. We construct the potential function Ψ using a greedy algorithm: start with the highest scoring triangle in the overlap graph and iteratively add vertices that increase the overall weight of coherent edges by the maximum amount possible at each step. Although the weight of the coherent edge-set returned by this procedure is not guaranteed to be maximal, it is likely to be so after an adequate threshold is imposed on β to select which edges to retain in the overlap graph.

Once a coherent edge-set E is found, one can construct another coherent edge-set from the edges symmetric to those in E . If the first of these corresponded to the alignment of the prefix masses, then the other would correspond to the alignment of the suffix masses (and vice versa).

Multiple Alignment of MS/MS Spectra. A set of coherent edges defines a multiple alignment (assembly) of PRM spectra. In this context, a multiple alignment is thus defined as a pair $A = (\mathcal{P}, \Psi)$ where $\mathcal{P} = \{P_1, \dots, P_n\}$ denotes a set of PRM spectra and $\Psi = \{\psi_1, \dots, \psi_n\}$ denotes the potentials (relative positions) of the PRM spectra. Then, scoring the individual PRMs over a multiple alignment A is very similar to what was described before for scoring over a cluster—for a putative PRM t , score it in each overlapped spectrum and set its consensus score $w(t, A)$ to the sum of the obtained per-spectrum PRM scores. The only difference is that in this case there are different starting positions Ψ

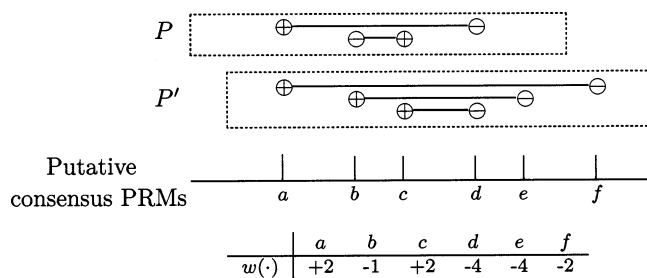


Figure 9. Putative consensus PRM scores for two aligned PRM spectra P (2 twin PRM pairs) and P' (3 twin PRM pairs) with fixed twin PRM orientations.

that need to be taken into account; scoring a PRM t over a multiple alignment $A = (\{P_1, \dots, P_n\}, \{\psi_1, \dots, \psi_n\})$ then becomes

$$w(t, A) = \sum_{t-\psi_i \in P_i} w(t - \psi_i, P_i)$$

Thus, a consensus PRM spectrum P for a multiple alignment A could be defined as the set $P = \{t: w(t, A) > 0\}$ —the set of all PRMs with a positive summed score. Although this is a reasonable first approach, it is sometimes the case that MS/MS peaks corresponding to neutral losses also generate PRMs in the aligned PRM spectra and could, therefore, also generate PRMs in the consensus PRM spectrum. This effect is minimized by requiring a minimum 57-Da distance between PRMs in a consensus (sparse subset).

A consensus PRM spectrum for a multiple alignment A is a sparse subset of P , where $P = \{t: w(t, A) > 0\}$ is the set of all PRMs with a positive score over A .

Avoiding Double-Counting in the Consensus PRM Spectrum. In a correct multiple alignment of ideal PRM spectra, it would be likely (although not certain) that only same-type PRMs would match; i.e., either all matched PRMs would be prefix masses or all would be suffix masses. In reality, this is not the case for two reasons. The first is that MS/MS spectrum peaks from neutral losses may also generate PRMs in a PRM spectrum—this was already minimized above by the definition of consensus as a sparse set. The second reason is that sometimes, due to random chance or local similarities, twin PRMs will both match other PRMs in a multiple alignment—if both were used in the consensus spectrum, the same MS/MS peaks would be counted twice (as described

under Prefix Residue Mass Spectra, twin PRMs are generated and scored by the same MS/MS spectrum peaks). To avoid this, let \oplus be the positive score of a correct PRM and let \ominus be the negative score of a PRM with no supporting MS/MS spectrum peaks. We define an *orientation* for every pair of twin PRMs (p, q) as orientation \oplus/\ominus when p is the prefix mass and q is its twin PRM; represented as

$$p\oplus-\ominus q$$

and orientation \ominus/\oplus when q is the prefix mass and p is its twin PRM, represented as

$$p\ominus-\oplus q$$

The score of the consensus PRMs is computed in exactly the same way—the sum of the matched PRM scores. Figure 9 illustrates this for a pair of PRM spectra with given twin PRM orientations and considering $\oplus = +1$ and $\ominus = -2$. The observed weights of the putative PRMs define a consensus PRM spectrum C with PRMs at positions $\{a, c\}$ and $w(C) = w(a) + w(c) = 4$.

Our problem then becomes the following: given a multiple alignment find a set of twin PRM orientations that yield a consensus PRM spectrum of maximal weight—*maximal oriented consensus problem*. In practice, this problem can be solved using a greedy approach—consider a multiple alignment A with unknown twin PRM orientations and assume that the score of nonoriented PRMs is given by \oplus . Then proceed as follows: (1) Select the putative PRM t with the highest aggregate score $w(t, A)$ (as described above) and assign it to the consensus PRM spectrum. (2) Mark the PRMs matching t in the multiple alignment as \oplus and their corresponding twins as \ominus . (3) Repeat from 1 until all aggregate scores are negative.

Step 2 above guarantees that there is no double counting of MS/MS spectrum peaks—whenever a PRM is selected as part of the consensus PRM spectrum, its twin PRM is marked as \ominus and thus will not contribute positively to the score of any other consensus PRM.

The preferential match of same-type PRMs (all prefix or all suffix) in a multiple alignment leads to the selective retention of same-type PRMs (as can be seen in Figures 10 and 11)—the mean percentage of PRM spectrum scores assigned to same-type PRMs is 95%. Even more interesting, these PRMs tend to form very clear

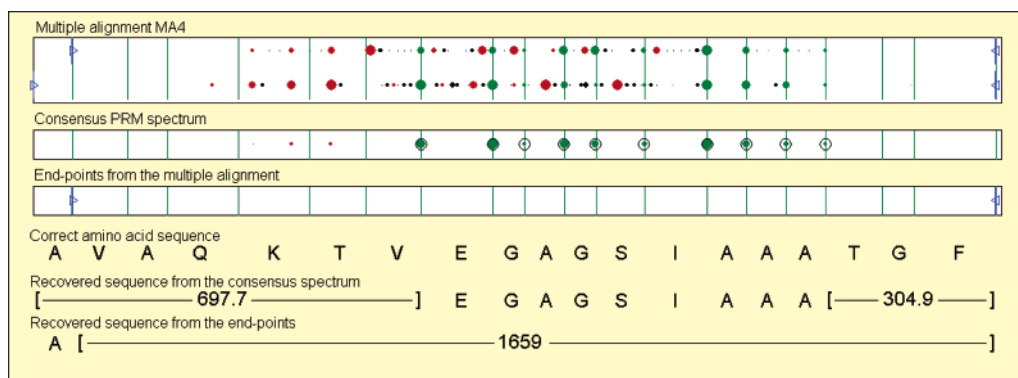


Figure 10. Resulting interpretation of the assembled PRM spectra in multiple alignment MA4 (Figure 12). In this case, unlike those shown in Figure 11, there is no match between internal PRMs in the consensus spectrum and the end points. As such, the end points only contribute that there is an alanine either at the start (correct answer) or at the end of the peptide but not its exact location.

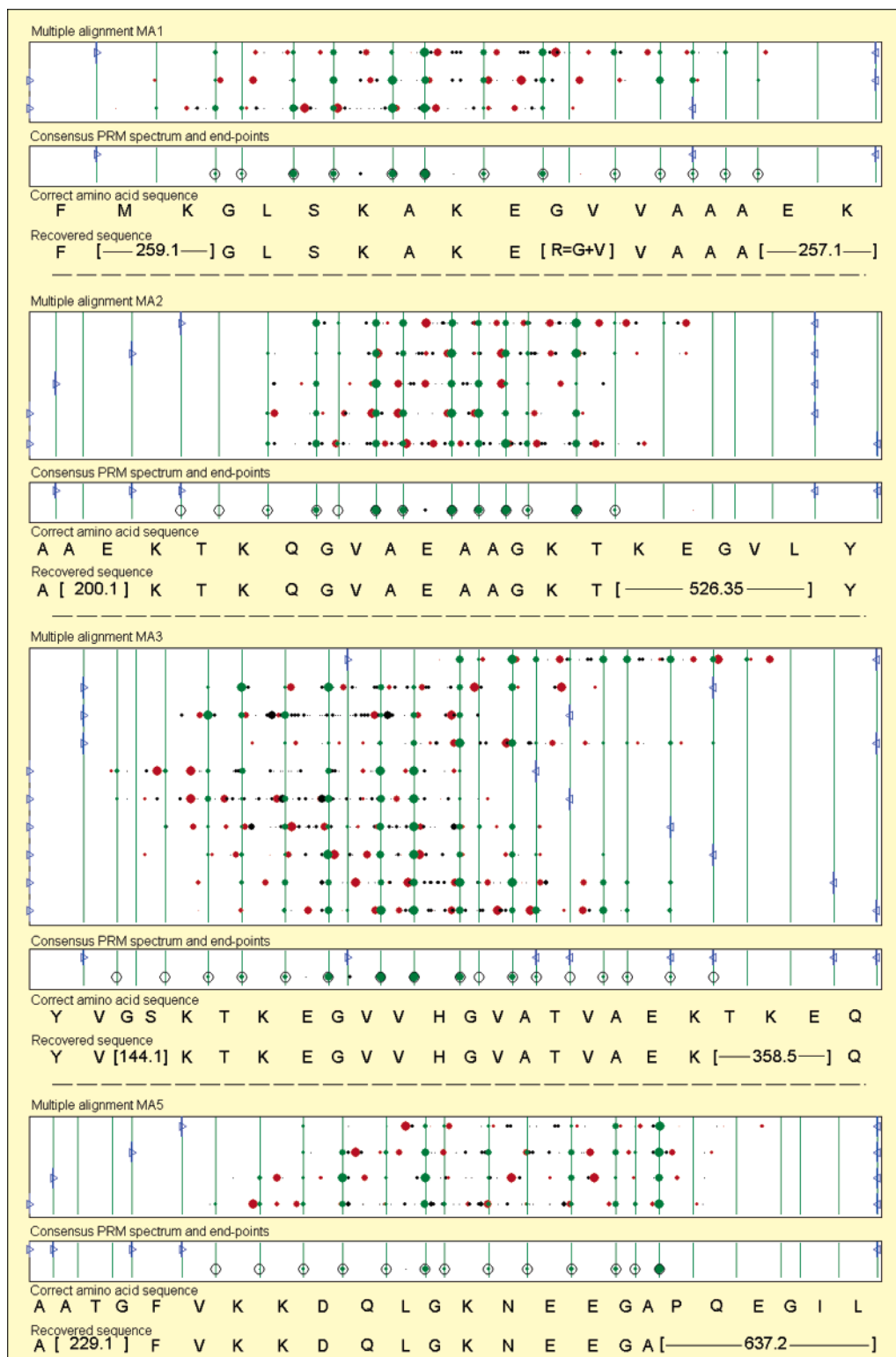


Figure 11. Resulting interpretation of the assembled PRM spectra in multiple alignments MA1, MA2, MA3 and, MA5.

ladders where the sequential mass differences correspond to amino acid masses, turning the de novo interpretation of the PRM spectra into a simple problem.

EXPERIMENTAL RESULTS

We evaluated our approach in a pilot study of a sample containing purified α -synuclein. The sample was digested using pepsin and a total of 2646 MS/MS spectra with precursor charge

+2/+3 were obtained using an ESI/ion trap mass spectrometer. After parent mass correction and precursor charge selection, we retained 1748 of these MS/MS spectra, all believed to be of charge +2. We chose not to include MS/MS spectra with precursor charge +3 due to the higher probability of charge +2 ion types (e.g., b^3), which would not align with the predominant single charge ion types in MS/MS spectra with parent charge +2.

Table 4

clusters	no. of spectra	spectra from α -synuclein	% correct
all	236	183	77.5
top 29 clusters	201	183	91.0
bottom 10 clusters	35	0	0

In the absence of an expert-annotated data set, we annotated the MS/MS spectra using Sequest. The 1748 MS/MS spectra were searched against a database of 10 000 protein sequences randomly selected from the NCBI database plus our sequence for α -synuclein. Sequest was configured to allow for a peptide precursor mass tolerance of 2 Da, spectrum peak tolerance of 0.5 Da, and nonspecific enzymatic digestion. This procedure identified 303 MS/MS spectra as peptides from α -synuclein (by considering the top peptide assignment only), which corresponds to a rate of 17% positive IDs or correct spectra. Once again we stress that these annotations and database search results are not used by our method in any way; these are used only to evaluate the quality of the results. This annotation strategy is of course biased toward what Sequest could do on a set of MS/MS spectra from a non-trypsin-specific digestion, but in the absence of an adequate and curated data set, it is a reasonable approximation to the true performance of our method.

In the clustering phase, match scores were computed over the set of 1748 spectra for every pair of PRM spectra with an absolute parent mass difference not larger than 2 Da. Around 83% of the spectra were matched to at least one other spectrum resulting in 236 spectra being retained in the obtained 39 clusters (the remaining spectra did not meet the clustering criteria): As can be seen from Table 4, most of the "incorrect" spectra retained were concentrated in 10 small clusters, which were later ignored in the alignment and assembly phases—the consensus PRM spectra obtained from these 10 clusters did not align to any other PRM spectrum.

The 39 clusters obtained from the clustering phase produced 39 consensus PRM spectra, which were then aligned using our pairwise alignment procedure as described under Aligning MS/MS Spectra. After adequately thresholding the alignment quality score β , we retained 114 relative shifts, all from correct alignments between α -synuclein spectra. These pairwise alignments defined five connected components in the overlap graph with consensus spectra and interpretations as shown in Figures 10 and 11. As shown in these, we were able to accurately recover large portions of the overlapped peptide regions. Another major advantage of our approach is also shown—the differences between the shifts (right-pointing triangles) and the parent masses (left-pointing triangles) of the aligned PRM spectra also correspond to amino acid masses. This fact allows us to reconstruct the amino acid sequences near the ends of the consensus PRM spectra even when absolutely no MS/MS spectrum contains any peaks for these fragments—we call these *end-point sequences*. In four out of five cases (Figure 11) at least one end point matches an internal PRM in the consensus PRM spectrum, either directly or by looking for PRMs at valid amino acid mass distances. In the single occasion where this is not the case (Figure 10), the end-point sequence yields additional peptide sequence information but the

orientation is not known—the shift of 71 Da only indicates that the peptide either starts (correct answer) or ends with alanine.

Figure 12 illustrates the position of the retained MS/MS spectra relative to the α -synuclein protein sequence. Boxes MA1–MA5 contain spectra participating in multiple alignments, and boxes C1–C3 contain spectra that clustered together but did not successfully align to other spectra. The recovered amino acid sequences are shown together in Figure 13—the identified sequence blocks (multiple alignments MA1–MA5 and clusters C1–C3) cover 90% of the whole protein and accurately recover 60% of the whole amino acid content.

The coverage gap near the end of the protein sequence is not caused by our method but rather a consequence of a very low MS/MS spectrum coverage in that specific area, observed even when using Sequest to search the database with the correct protein sequence. This was an area of high enzymatic cleavage by pepsin, which did not generate enough MS/MS spectra from peptides covering this and adjacent areas, and also, the two proline amino acids near the center of the gap promote the absence of valuable MS/MS peaks when attempting clustering or alignment in this region.

DISCUSSION

Since the onset of the application of tandem mass spectrometry to the analysis of peptide sequences,^{28–30} many techniques have been proposed to interpret peptide spectra; recent reviews of the area are available in refs 31 and 32. One of these approaches set the basis for the popular tool Sequest¹⁶ by showing how to search a protein sequence database to identify a peptide with a given MS/MS spectrum. Several refinements to the initial approach have been developed on how to score the matches between an MS/MS spectrum and a peptide from a database^{17,22,33–38} and on how to assess the reliability of the putative peptide assignments.^{39–41} Recent developments also include database search techniques to identify MS/MS spectra from peptides with posttranslational modifications.^{10,18,42–45} Another approach, *de novo* peptide interpretation, attempts to find the amino acid sequence that best explains the MS/MS spectrum. Although *de novo* peptide

- (28) Hunt, D.; Bone, W.; Shabanowitz, J.; Rhodes, J.; Ballard, J. *Anal. Chem.* **1981**, *53*, 1704–1706.
- (29) Aberth, W.; Straub, K.; Burlingame, A. *Anal. Chem.* **1981**, *54*, 2029–2034.
- (30) Hunt, D. F.; Yates, J. R.; Shabanowitz, J.; Winston, S.; Hauer, C. R. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83* (17), 6233–6237.
- (31) Aebersold, R.; Mann, M. *Nature* **2003**, *422* (6928), 198–207.
- (32) Mann, M.; Jensen, O. N. *Nat. Biotechnol.* **2003**, *21* (3), 255–261.
- (33) Yates, J. R.; Eng, J. K.; McCormack, A. L.; Schieltz, D. *Anal. Chem.* **1995**, *67* (8), 1426–1436.
- (34) Bafna, V.; Edwards, N. *Bioinformatics* **2001**, *17* (Suppl 1), 13–21.
- (35) Sadygov, R. G.; Yates, J. R. *Anal. Chem.* **2003**, *75* (15), 3792–3798.
- (36) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20* (18), 3551–3567.
- (37) Creasy, D. M.; Cottrell, J. S. *Proteomics* **2002**, *2* (10), 1426–1434.
- (38) Lu, B.; Chen, T. *Bioinformatics* **2003**, *19* (Suppl 2), 113–113.
- (39) MacCoss, M. J.; Wu, C. C.; Yates, J. R. *Anal. Chem.* **2002**, *74* (21), 5593–5599.
- (40) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74* (20), 5383–5392.
- (41) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2003**, *75* (17), 4646–4658.
- (42) Pevzner, P. A.; Dancik, V.; Tang, C. L. *J. Comput. Biol.* **2000**, *7* (6), 777–787.
- (43) Pevzner, P. A.; Mulyukov, Z.; Dancik, V.; Tang, C. L. *Genome Res.* **2001**, *11* (2), 290–299.
- (44) Liebler, D. C.; Hansen, B. T.; Davey, S. W.; Tiscareno, L.; Mason, D. E. *Anal. Chem.* **2002**, *74* (1), 203–210.

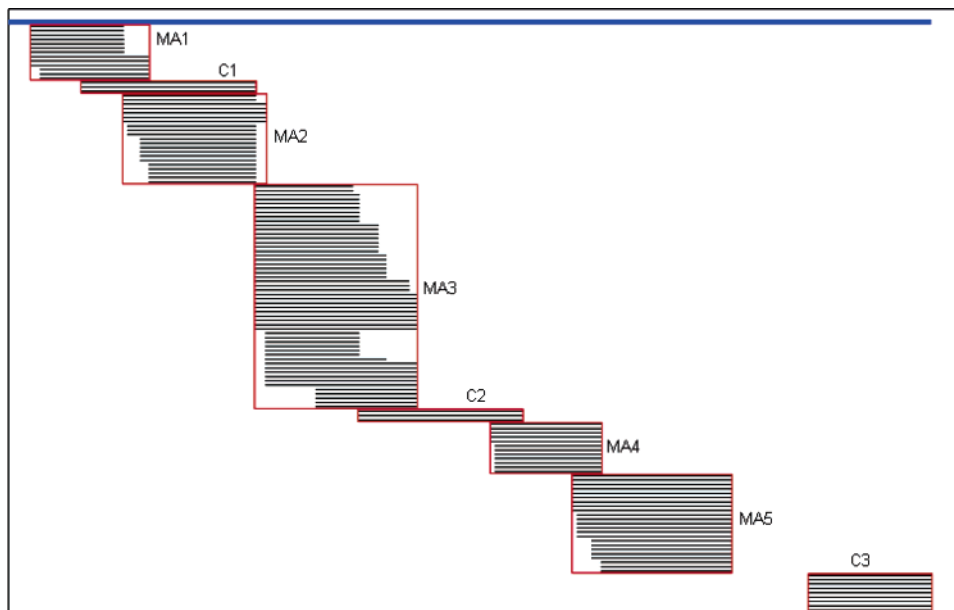


Figure 12. α -Synuclein spectra clustered (C1–C3) and assembled (MA1–MA5). The blue line at the top represents the complete protein sequence.

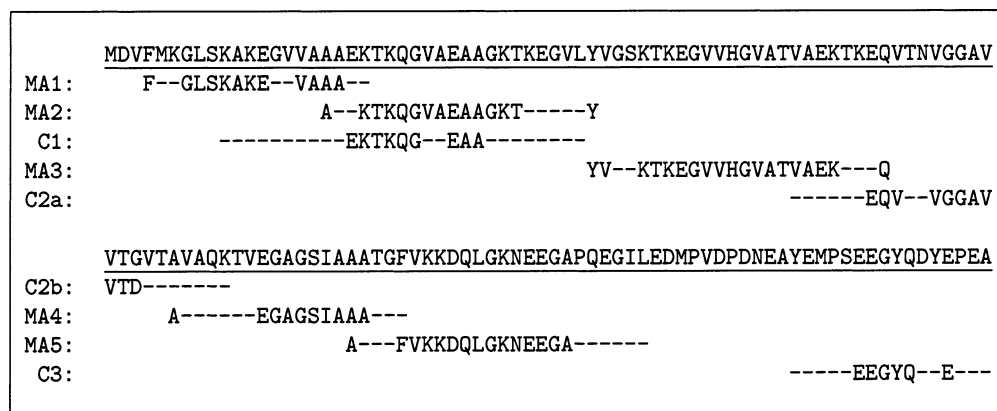


Figure 13. Recovered portions of the α -synuclein protein sequence. The recovered sequences are shown under the correct (underlined) protein sequence; dashes indicate portions where the mass intervals are known but the exact amino acid sequences are not. The only incorrectly identified amino acid is shown in red. The interpretation of cluster C2 was split into two lines as C2a and C2b; the full interpretation is the concatenation of these two.

sequencing can be viewed as a search in the virtual database of all possible peptides, the algorithms for de novo peptide sequencing are very different from those used in peptide identification via database search. Some examples of de novo peptide sequencing algorithms include Sherenga,¹¹ SeqMS,⁴⁶ Lutefisk,^{47,48} Peaks,⁴⁹ and approaches developed by Chen et al.²⁶ and Bafna and Edwards.²⁷

Tandem mass spectrometry combined with database search tools is a successful approach to peptide identification. When the quality of the MS/MS spectra is reasonable and the corresponding

peptide sequences are in the database, these tools generally find the correct MS/MS interpretation. Unfortunately, this is not always the case due to several factors: alternative splicing variants, missing genes, or simply the absence of any protein sequence information for the context of the experiment. The latter is the case when studying proteins from snake and scorpion venom,^{50–54} whose sequences had to be determined by Edman degradation. Database search tools also face severe difficulties when attempting

(45) Searle, B. C.; Dasari, S.; Turner, M.; Reddy, A. P.; Choi, D.; Wilmarth, P. A.; McCormack, A. L.; David, L. L.; Nagalla, S. R. *Anal. Chem.* **2004**, *76* (8), 2220–2230.
(46) Fernandez-de Cossio, J.; Gonzalez, J.; Satomi, Y.; Shima, T.; Okumura, N.; Besada, V.; Betancourt, L.; Padron, G.; Shimonishi, Y.; Takao, T. *Electrophoresis* **2000**, *21* (9), 1694–1699.
(47) Johnson, R. S.; Taylor, J. A. *Methods Mol. Biol.* **2000**, *146*, 41–61.
(48) Johnson, R. S.; Taylor, J. A. *Mol. Biotechnol.* **2002**, *22* (3), 301–315.
(49) Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G. *Rapid Commun. Mass Spectrom.* **2003**, *17* (20), 2337–2342.

(50) D'Suze, G.; Sevcik, C.; Corona, M.; Zamudio, F. Z.; Batista, C. V.; Coronas, F. I.; Possani, L. D. *Toxicon* **2004**, *43* (3), 263–272.
(51) Ogawa, Y.; Yanoshita, R.; Kuch, U.; Samejima, Y.; Mebs, D. *Toxicon* **2004**, *43* (7), 855–858.
(52) Xu, C. Q.; He, L. L.; Brône, B.; Martin-Eauclaire, M. F.; Van Kerkhove, E.; Zhou, Z.; Chi, C. W. *Toxicon* **2004**, *43* (8), 961–971.
(53) Nirthanan, S.; Charpantier, E.; Gopalakrishnakone, P.; Gwee, M. C.; Khoo, H. E.; Cheah, L. S.; Bertrand, D.; Kini, R. M. *J. Biol. Chem.* **2002**, *277* (20), 17811–17820.
(54) Scarborough, R. M.; Rose, J. W.; Naughton, M. A.; Phillips, D. R.; Nannizzi, L.; Arfsten, A.; Campbell, A. M.; Charo, I. F. *J. Biol. Chem.* **1993**, *268* (2), 1058–1065.

identification of MS/MS spectra from posttranslationally modified and mutated peptides. The requirement to consider every expected modification (or set of modifications) over every possible candidate peptide not only makes the process much slower but also reduces the reliability of interpretations due to the highly expanded pool of peptide candidates. The latter is also a problem faced by de novo analysis tools, which efficiently select the best peptide to annotate a spectrum from the set of all possible peptides with the same precursor mass. These tools eliminate the requirement for a database of protein sequences and, when used on high-quality data, generally output correct amino acid sequences explaining large portions of the spectrum (the remaining portions, usually near the start and end of the spectrum, tend to have very few or no peaks). Another serious difficulty faced by current de novo analysis tools is that the low signal-to-noise ratio in most MS/MS spectra leads to the generation of not one but several high-scoring peptide candidates.

The method presented in this paper builds on strengths from previous approaches to generate larger and more reliable peptide sequences without requiring an existing database of protein sequences. In our approach, spectra are compared against each other (similar to the comparison of experimental spectra to theoretical spectra in database search) to detect repeated MS/MS spectra from the same peptide, which are then used to effectively increase the signal-to-noise ratio. The same principle is also applied to detect partial overlaps between spectra and assemble them into multiple alignments where the evidence for real fragment masses becomes overwhelming when compared to that available in single spectra. Furthermore, the multiple alignments themselves provide additional valuable information in that the end points of the aligned spectra must necessarily correspond to interresidue points in the protein sequence and provide, for the first time, a way to recover sequence information where absolutely no MS/MS spectrum peaks are available. Altogether, we build on the ideas previously applied to DNA sequencing to significantly improve the de novo analysis of amino acid sequences and take it from single peptide sequencing to the level of protein sequencing.

Moreover, our approach is directly applicable to sets of MS/MS spectra from posttranslationally modified proteins. Because we make no assumptions on the set of residue masses (other than a minimum residue mass of 57 Da), the same procedure can be used to seamlessly assemble spectra from modified peptides and directly determine the modified protein sequence (work in progress). Related work by MacCoss et al.¹⁰ has shown that the analysis of partially overlapping peptides provides valuable evi-

dence toward confirming the presence of posttranslational modifications. Along the same lines of reasoning, even when complete protein coverage is not available, our method can be used to increase the confidence of de novo interpretations by supporting the peptide sequences reconstruction with several partially overlapping spectra.

From an experimental perspective, our approach does not require any new developments or significant changes to the currently known protocols; the single difference is that instead of using only trypsin as a digestion enzyme (for which database search tools are specifically tailored), nonspecific enzymes (or sets of enzymes with different specificity) should be used. As Woods and co-workers have shown,^{5-7,9} the generation of rich peptide ladders is feasible and within reach of readily available technology. The fact that our results were produced using data from ESI/ion trap mass spectrometers further reinforces this point: although higher mass accuracy instruments such as MALDI-qTOF and MALDI-TOF/TOF should greatly enhance the quality of the sequence reconstruction, they are not required for our method to be applicable.

The major difficulty faced by our method was the quality of the experimental MS/MS spectra. This was circumvented by the application of clustering and filtering techniques but at the cost of reduced protein sequence coverage. The future availability of larger data sets generated by adequate experimental protocols will allow us to better estimate both the necessary peptide coverage for complete protein sequencing and rigorous thresholds for statistically significant matches. Also, although the occurrence of long repeats in the protein sequence could be an issue, this does not seem to be a frequent event. Most repeated subsequences tend to be very short and completely covered by several longer peptides and thus do not significantly affect our approach.

ACKNOWLEDGMENT

The authors thank Virgil Woods, Chris Gessner, and Dennis Pantazatos for providing the set of experimental MS/MS spectra from α -synuclein used to develop and test our algorithm. This project was supported by NIH Grant NIGMS 1-R01-RR16522.

SUPPORTING INFORMATION AVAILABLE

Additional information as noted in text. This material is available free of charge via the Internet at <http://pubs.acs.org>.

Received for review July 23, 2004. Accepted September 8, 2004.

AC0489162