Multi-spectra peptide sequencing and its applications to multistage mass spectrometry

Nuno Bandeira1,*, Jesper V. Olsen2, Matthias Mann2 and Pavel A. Pevzner2
1Department of Computer Science and Engineering, University of California, San Diego, USA and 2Max-Planck Institute for Biochemistry, Germany

ABSTRACT
Despite a recent surge of interest in database-independent peptide identifications, accurate de novo peptide sequencing remains an elusive goal. While the recently introduced spectral network approach resulted in accurate peptide sequencing in low-complexity samples, its success depends on the chance of presence of spectra from overlapping peptides. On the other hand, while multistage mass spectrometry (collecting multiple MS2 spectra from each MS1 spectrum) can be applied to all spectra in a complex sample, there are currently no software tools for de novo peptide sequencing by multistage mass spectrometry. We describe a rigorous probabilistic framework for analyzing spectra of overlapping peptides and show how to apply it for multistage mass spectrometry. Our software results in both accurate de novo peptide sequencing from multistage mass spectra (despite the inferior quality of MS2 spectra) and improved interpretation of spectral networks. We further study the problem of de novo peptide sequencing with accurate parent mass (but inaccurate fragment masses), the protocol that may soon become the dominant mode of spectral acquisition. Most existing peptide sequencing algorithms (based on the spectrum graph approach) do not track the accurate parent mass and are thus not equipped for solving this problem. We describe a de novo peptide sequencing algorithm aimed at this experimental protocol and show that it improves the sequencing accuracy on both tandem and multistage mass spectrometry.

Availability: The open-source implementation of our software is available at http://proteomics.bioprojects.org.
Contact: bandeira@ucsd.edu
Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION
The coupling of tandem mass spectrometry (MS2) with database search tools (Perkins et al., 1999; Tanner et al., 2005; Yates et al., 1995) is the enabling core behind high-throughput protein identification ( Aebersold and Mann, 2003). Unfortunately, this successful strategy is not applicable to unknown protein sequences. Particularly important examples of value derived from initially successful strategy is not applicable to unknown protein sequences. Antibodies illustrate the scenario where the universe of possible protein sequences is very large and constantly altered by recombination and somatic hypermutation (Maizels, 2005). Additionally, drugs derived from venom proteins exemplify benefits of exploring the viable proteins already probed by natural biodiversity. Another recently emerged application of de novo peptide sequencing is proteinogenetics, using mass spectra to predict new genes and alternatively spliced variants (Edwards, 2007; Gupta et al., 2007; Tanner et al., 2007). Proteinogenetics studies would greatly benefit from searching 6-frame translations of entire genomes that amounts to roughly 6 billion amino acids for the mammalian genomes. The database search approaches are impractical for such large searches while de novo-based approaches to protein identifications (e.g. Alves and Yu, 2005) are not seriously affected by the database size.

MS-based studies of unknown proteins rely on de novo peptide sequencing techniques that attempt to recover the peptides directly from the spectra. Unfortunately, despite decades of research, accurate de novo peptide sequencing remains an elusive goal with only 30–40% of all peptides (depending on the sample and peptide size) reconstructed correctly (Pevtsov et al., 2006). While algorithms have been developed to find a best-‘scoring’ peptide for a given spectrum (Bafna and Edwards, 2003; Chen et al., 2001; Dancik et al., 1999; Fischer et al., 2005; Frank and Pevzner, 2005; Ma et al., 2003), their sequencing accuracy is still affected by incomplete fragmentation, noise and ambiguity in ion-type assignments. In abundant, low-complexity samples, these difficulties may be attenuated by generating overlapping peptides and combining the resulting MS2 spectra to yield higher accuracy de novo sequences—examples include 18O/16O labeling (Shevchenko et al., 1997) or Shotgun Protein Sequencing (Bandeira et al., 2004, 2007a). While Shotgun Protein Sequencing was shown to work well for low-complexity samples it remains unclear how it will scale up for complex samples. Multistage MS (e.g. generating multiple MS3 spectra for each MS2 spectrum) could be an ideal method to address the potential limitations of Shotgun Protein Sequencing in a controlled and even experiment-driven fashion. However, while various groups recently demonstrated the advantages of multistage MS for database search (Kalkum et al., 2003; Olsen and Mann, 2004; Ulintz et al., 2008), there are currently no software tools for de novo interpretation of multistage MS data. As a result, the multistage MS data are still manually interpreted, not unlike tandem MS in mid-1990s, before the first de novo peptide sequencing tools were developed.

Given a spectrum s, de novo peptide sequencing algorithms attempt to find a peptide π that best ‘explains’ the spectrum (e.g. maximizes the probability of generating the spectrum s). In multi-spectra peptide sequencing, one attempts to find a peptide π that best explains multiple spectra s1,...,sk at once. Multistage MS is...
Multi-spectra peptide sequencing

Table 1. Spectrum ion statistics for the yeast dataset

| Type of spectra | Number of spectra | Intensity (%) | Number of peaks | \( p(b) \) | \( p(y) \) |
|-----------------|------------------|--------------|----------------|----------|----------|
| a) MS/MS (MS\(^2\)) | 890 | b 13 | y 30 | Satellite 20 | Unexplained 37 | 542 | 0.81 | 0.84 |
| MS/MS/MS (MS\(^3\)) | 4447 | | | | | | | |
| from b-ions | 1039 | b 11 | y 17 | Satellite 23 | Unexplained 49 | 62 | 0.41 | 0.51 |
| from y-ions | 2592 | | | | | | | |
| b) Scored MS\(^2\) | 890 | b 28 | y 28 | Satellite 23 | Unexplained 31 | 80 | 0.81 | 0.81 |
| Scored MS\(^3\) | 4447 | | | | | | | |
| from b-ions | 1039 | b 14 | y 3 | Satellite 22 | Unexplained 61 | 56 | 0.30 | 0.12 |
| from y-ions | 2592 | | | | | | | |
| c) Merged (MS\(^2\) + MS\(^3\)) | 602 | b 66 | y 8 | Satellite 7 | Unexplained 19 | 83 | 0.89 | 0.77 |
| d) Merged + sequenced | 602 | b 92 | y 1 | Satellite 1 | Unexplained 6 | 11 | 0.82 | 0.01 |

a) Ion statistics for RAW spectra. b) Ion statistics after replacing each peak’s raw intensity with a likelihood score. (Frank and Pevzner, 2005). c) Ion statistics of the consensus spectra obtained by merging each MS\(^2\) spectrum with its dependent MS\(^3\) spectra (after scoring). This category has less spectra (602) than the others (890) mainly because not all MS\(^3\) spectra had at least one ‘usable’ MS\(^3\) spectrum (details in main text). d) Ion statistics for the peaks selected by de novo sequencing on the merged spectra.

as an example of the multi-spectra peptide sequencing problem. In certain aspects, interpretation of multistage MS data is more challenging than interpretation of multiple tandem mass spectra (as in Bandeira et al., 2007a,b) since MS\(^3\) spectra typically have lower quality than MS\(^2\) spectra. The manual usage of MS\(^3\) spectra as an aid to de novo sequencing dates back to a decade ago (Lin and Glish, 1998). To the best of our knowledge, the only automated approach to multistage MS is a heuristic approach proposed by Zhang and McElvain (2000). However, the sequencing accuracy of this approach remains unknown since Zhang and McElvain were limited to only 42 sets of multistage spectra and the accuracy gains from the proposed heuristics were described qualitatively rather than quantitatively. Moreover, this approach has not enthused further MS\(^2\)/MS\(^3\) sequencing efforts because no implementation is available. As an addition (or alternative) to overlapping spectra, we also proposed an efficient dynamic programming algorithm for de novo peptide sequencing with accurate peptide masses and evaluate the resulting contribution to de novo sequencing accuracy.

2 METHODS

Datasets The yeast dataset was acquired from a trypsin digestion of a yeast whole-cell lysate on a Thermo LTQ-FT instrument configured for the acquisition of up to five MS\(^2\) spectra from each MS\(^2\) spectrum with a doubly charged precursor; precursor masses for the MS\(^3\) spectra were restricted to masses higher than the MS\(^2\) precursor mass. As a result, we obtained a total of 15770 dependent MS\(^2\) spectra. These spectra were searched against a yeast database using InsPecT (details in main text). We also tested it on 10 517 MS\(^2\) spectra.1 In general, we observed that MS\(^3\) spectra tend to have less explained intensity and less b/y-ion peaks than MS\(^2\) spectra (details shown in Table 1). The shewanella dataset was acquired on Thermo LTQ instruments over many LC/MS/MS runs with different multi-dimensional separation techniques (Gupta et al., 2007). For the purpose of this study we arbitrarily selected 34 326 MS\(^2\) spectra from different peptides, resulting in 10 517 MS\(^2\) spectra with between one and five other MS\(^2\) spectra from prefix/suffix peptides. From a computational perspective, this shewanella MS\(^2\) dataset mimics the yeast MS\(^2\)/MS\(^3\) dataset with unrealistically high quality of MS\(^3\) spectra and thus allows one to establish an upper bound on the performance of our multistage MS approach.

Spectrum identification Similar to the protocol used for the identification of spectra in the shewanella dataset (Gupta et al., 2007), InsPecT (Tanner et al., 2005) was used to search a database containing 7517 Saccharomyces cerevisiae protein sequences (SwissPROT, October 8, 2006) and sequences of common contaminant proteins. The database additionally contained 7517 decoy protein sequences used to enforce the selected 5% false-discovery rate. InsPecT was configured to allow for a 0.5 Da fragment mass tolerance; the high accuracy of the experimental precursor masses was used to confirm identifications rather than to restrict the

\(^{a}\)A total of 1039 MS\(^1\) spectra were annotated as prefix fragments (b-ions), 2592 MS\(^3\) spectra were annotated as suffix fragments (y-ions) and 320 spectra were annotated as prefix/suffix fragments after loss of H\(_2\)O or NH\(_3\).
space of possible peptides. The set of allowed modifications was oxidation (M), phosphorylation (S, T and Y), acetylation (N-term), deamidation (Q) and

Given a spectrum \( s \) and a set of strings \( \Pi \), we are interested in solving the optimization problem of finding \( \max_{\pi \in \Pi} \Pr(s | \pi) \). Below we focus on the sets \( \Pi \) that are relevant in the context of tandem MS. Let \( V = \{1, \ldots, n\} \) and \( G(V, E) \) be a topological ordering of a Directed Acyclic Graph (DAG) such that \( i < j \) for every directed edge \((i, j)\) in \( E \). Every path from 1 to \( n \) in graph \( G \) corresponds to a G-peptide \( \pi = \pi_1 \ldots \pi_n \) such that \( \pi_i \) is 1 if vertex \( i \) belongs to the path. We are interested in the following peptide sequencing (PS) problem.

**PS problem** Given a spectrum \( s \) and a DAG \( G \), find a G-peptide \( \pi \) maximizing \( \Pr(s | \pi) \) over all G-peptides.

We impose no restrictions on the graph \( G(V, E) \), but in practical applications it is usually assumed that \((i, j) \in E \) iff \( i < j \) equals the integer mass of an amino acid. Such graphs are referred to as spectrum graphs (Bartels, 1990; in MS, they usually encode all peptides of a given parent mass \( m \)).

The relation between this abstract Boolean strings model and de novo peptide sequencing is straightforward. In reality, an MS/MS spectrum can be represented as a string of ions (peak present) and zeros (no peak present), with a 0/1 for every consecutive 1 Da interval. Similarly, sequences of amino acid masses (peptides) can also be represented as strings of zeros and ones. Every amino acid can be represented as a string of \( a-1 \) zeros followed by a single one, where \( a \) is the integer amino acid mass. Then, a peptide is simply a concatenation of Boolean strings corresponding to its amino acids. In this context, \( \theta \approx 0.05 \) (probability of observing a noise peak) and \( \rho \approx 0.7 \) (probability of observing a b-ion) represent typical values of \( \theta \) and \( \rho \) for ion-trap MS/MS spectra (see table in Fig. 1A). This somewhat simplistic Boolean string model can be modified for any mass resolution, peptide-fragmentation rules and peak intensities (Baliga and Edwards, 2003; Chen et al., 2001; Frank and Pevzner, 2005). Moreover making this model more realistic typically does not affect the algorithmic solution. In particular, (Bandeira et al., 2007b) recently showed how spectral alignment can be used to separate between b/y-ions and thus generate strings as in this model.

The PS Problem has an easy solution first described by (Dancik et al., 1999). We will find it convenient to cast the approach in (Dancik et al., 1999) as an application of the Viterbi algorithm (Durbin et al., 1999; Viterbi, 1967) in an appropriately constructed hidden Markov model (HMM) \( G \). Figure 1 shows a graph \( G \) with every edge \((i, j)\) substituted by a path with new edge \( j-i \) traversed vertices that start at \( i \) and end at \( j \). The resulting graph with vertex set \( V \cup T \) (where \( V = \{1, \ldots, n\} \) and \( T \) is the set of all traversal vertices) represents the hidden states of the HMM and all possible transitions between the states. The emission probabilities of the HMM are defined by the matrix \( \Pr(s | \pi) \) with \( \Pr(s_i = 1 | T) \) for \( T \) states and \( \Pr(s_i = 1 | \pi) \) for all other states (see Fig. 1A). The transition probabilities of all edges in
this HMM are defined to be 1.5 Finding an optimal path in this HMM is a straightforward application of the Viterbi algorithm. The model above does not capture the fact that MS/MS spectra represent both pre- and postx ions (b-ions series) and suffix ions (y-ions series). To reflect this we represent peptides as strings in 3-letter alphabet: 1 (theoretical y-cut) and 0 (no cut). Given a peptide \( \pi = \pi_1, \ldots, \pi_n \), we define its reverse as the peptide \( \pi = \pi_n, \ldots, \pi_1 \). We now redefine the probability of peptide \( \pi \) generating spectrum \( s \) as \( \text{Prob}(\pi|s) = \prod_i \text{Prob}(\pi_i|s_i) \), where \( \text{Prob}(\pi_i|s_i) \) is a 2 \( \times \) 3 matrix. In general, \( k \) ions can be modeled by \( k + 1 \) parameter columns in this matrix; note that using additional ion-types only changes the emission probabilities but not the construction of the HMM.

However, this and other formulations of the peptide sequencing problem encode a particular bias toward peptides that have b-ions \( b_i \) and y-ions \( y_i \) such that \( b_i \) and \( b_j \) add up to the peptide mass. In these cases, both ions use the same masses in the spectrum to artificially increase the score of the peptide but do so with conflicting ion type assignments (because \( b_i = y_j \) and \( b_j = y_i \)). Peptides that do not have such pairs of b-ions are referred to as anti-symmetric peptides and efficient algorithms are available to find the maximum scoring anti-symmetric peptide for any given spectrum (Bafna and Edwards, 2003; Chen et al., 2001). Nevertheless, ambiguous b/y-ion assignments remain one of the main sources of de novo sequencing errors. We show below how MS3 spectra can help resolve these ambiguities.

Although this model still essentially amounts to maximizing the weighted number of matched masses between a spectrum and a peptide, it already captures enough detail to allow us to describe the proposed extensions for combining MS3/MS5 spectra. In practice, this same framework is used to model more elaborate events to take into consideration the intensity of the peaks in the spectrum and to account for the presence/absence of other ion types (e.g. \( b - H_2\text{O} \) ions). As shown in Table 1, using more elaborate scoring from (Frank and Pevzner, 2005), one can replace raw intensities with peak assignments to the same spectrum peaks in the scoring function.

We now consider the following Multi-spectra peptide sequencing (MSPS) problem. **MSPS problem** Given spectra \( s^1, \ldots, s^k \) and a substring mapping \( \phi \), find a G-peptide \( \pi \) maximizing \( \text{Prob}(\pi|s^1, \ldots, s^k) = \prod_{i=1}^k \text{Prob}(\pi|s_i) \). It is easy to see that the HMM constructed for the PS Problem (described above) can also solve the MSPS problem. The only difference between the two HMMs resides in the types of symbols emitted by the hidden states. Let \( S(i) \) be the set of all spectrum values generated from \( \pi_i \) (defined by spectra \( s^1, \ldots, s^k \) and the mapping \( \phi \)). Then, each hidden state now emits a set of independent values \( S(i) = \{v_j\} \), with \( \pi_{ij} = \pi_i | v_j \) given by the table in Figure 1 as before: \( P(i|j) = 0 \) for all states and \( P(i|j) = 1 \) for all other states. Assuming all \( v_j \) are independent observations, the probability of observing any given set \( S(i) \) is \( \prod_{j=1}^m P(i|j) \). Once again, the MSPS problem can be solved by an anti-symmetric sequencing algorithm on \( \phi \) (Bafna and Edwards, 2003; Chen et al., 2001). We note that although multiple ion-type assignments can be avoided in the MS2 spectrum, the resulting peptide may be somewhat biased by conflicting ion-type assignments in the MS3 spectrum. While these conflicts could be readily avoided when only one overlapping spectrum is available (as described in Bandeira et al., 2007b), extending the algorithm to more overlapping spectra would lead to a significant computational burden unlikely to result in relevant gains in terms of sequencing accuracy.

A possible change to the MSPS problem would be to require the start/end positions of each spectrum as mandatory 1's in the returned peptide \( \pi \). This modified problem can also be solved using the strategy described above if we modify the HMM by removing all \( T \) states at the start/end positions of every spectrum. As such, every path through this modified HMM \( \phi \) will be forced to use the states corresponding to these start/end positions and thus results only in peptides that are consistent with the alignment of MS3 spectra.

**Aligning MS5/MS3 spectra** Under common experimental conditions, the highest intensity peaks in MS5 spectra typically correspond to y- and, to a lesser extent, b-ions. Another common source of high-intensity peaks are doubly charged ions. However, on MS2 from doubly charged peptide one can avoid doubly charged fragment ions by considering only peaks with

---

**Figure 1** Conceptual illustration of peak scoring on an MS2 spectrum (e.g. from PEPITIDE) aligned with its dependent MS3 spectra (e.g. S2 from PEPITIDE and S3 from EPITIDE). Since the correct alignment between MS5/MS3 spectra is not known a priori, our algorithm finds the highest scoring peptide over all possible alignments. M indicates the parent mass of the MS3 spectrum and complementary masses (e.g. b/y-ions whose masses add up to the corresponding parent mass) are connected by arcs under each spectrum. Colors indicate the sets of masses accounted for when scoring different peaks; the colored squares in the equations represent the summed scores of all peaks of the same color; violet peaks indicate the b-ions masses dictated by the parent masses of the MS3 spectra. Intrinsically, the score of a mass m summarizes all corroborating evidence for it being generated from a prex fragment. The red and blue peaks illustrate the contribution of the MS3 spectra to the separation of b/y-ions in the MS2 spectrum—intuitively, if m is a b-ion and M − m is not then the red peaks should be more prominent than the blue peaks.
the mass error incurred by modeling the peak mass difference \( k \) the maximum gapped-peptide score over all anti-symmetric gapped-peptides with a prefix ending at peak \( k \) explores all \( 2^b \) possible ion-type assignments for an MS2 spectrum with \( b \) ions in the MS3 precursor ion.

Given an MS3 spectrum from a particular peptide (e.g. PTIDE), the generation of an MS3 spectrum from one of its b-ions yields additional information about the corresponding prefix peptide (e.g. PTIDE). The converse reasoning applies to MS3 spectra from y-ions and suffix peptides (e.g. PTIDE). As such, determining the correct alignment between an MS3 spectrum and its parent MS2 spectrum essentially reduces to determining the b-ion type of the MS3 precursor ion.

A peptide is accepted with score \( S[u,v,k]\) if \( \delta_{u,v,k} : |k + m(v) - m(u) - a| \leq T \), where \( T \) is the chosen parent mass tolerance.

Fig. 3. Dynamic programming recursion for de novo peptide sequencing of a spectrum \( S \) with accurate parent mass \( M \). The variable \( S[u,v,k]\) represents the maximum gapped-peptide score over all anti-symmetric gapped-peptides with a prefix ending at peak \( u \) at peak \( v \) and \( b \) cumulative mass error \( k \); a gapped-peptide at position \( S[u,v,k]\) is defined as a (possibly incomplete) peptide consisting of a prefix peptide ending at peak \( u \), a gap of mass \( m(v) - m(u) \) and a suffix peptide starting at peak \( v \). Used notation: \( m(x) \) — mass of peak \( x \); \( g(x) \) — score of peak \( x \); \( A \) — set of amino acid masses; \( a \) is the mass error incurred by modeling the peak mass difference \( m(v) - m(u) \) with the amino acid mass \( a \). In reality, user-specified fragment mass tolerances are allowed when appropriate. Also, the set of amino acid masses \( A \) may include summed masses of pairs of amino acids (to allow for missing ions in the spectrum).

a mass higher than that of the precursor ion. By restricting the selection of MS3 precursor ions to this high-mass region our experimental setup (i) implicitly selected singly charged MS3 precursors from doubly charged MS2 precursors and (ii) biased toward MS3 spectra generated from b//y-ions. As such, determining the correct alignment between an MS3 spectrum and its parent MS2 spectrum essentially reduces to determining the b-ion type of the MS3 precursor ion.

In practical terms, the memory requirements for the third dimension spanned by \( k \) can be controlled by binning the cumulative mass errors at a chosen resolution \( r \). For typical values of \( r = 0.5 \) Da, \( r = 0.01 \) Da and \( T = 1 \) this results in a need for 100 times more memory than would have otherwise been needed. The memory requirements are on the order of \( O(n^t/r^t) \), where \( n \) is the number of peaks in the spectrum. Since \( n \) is usually \( \geq 100 \), these requirements can be easily met on any current desktop computer. We further note that the mass-error tolerances can easily be adjusted for each individual peak by varying the allowed range for \( k \). In particular, peaks derived from the parent masses of overlapping peptides may be set to much lower mass tolerances than peaks from fragment masses.

3 RESULTS

When analyzed in isolation, individual MS3 spectra are less useful than individual MS2 spectra. Lower amounts of substrate, low charge and a bias toward shorter peptides result in spectra with the mass array data structure (Mo et al., 2007), which defines mass bins with resolution \( r = 0.01 \) Da over the whole spectrum. However, we note that extending the mass-array approach to enforce the anti-symmetric condition would result in prohibitive-memory requirements on the order of \( O(nM^3/r^3) \), where \( M \) could be anywhere between 400 and 6000 (for triple-charged precursor ions).
also from the yeast and shewanella datasets were most similar, we might expect the lower average sequencing accuracy with 5 overlapping MS2 spectra from triply charged precursors still remains slightly inferior, it roughly approaches (or exceeds, when accurate parent masses are available) the sequencing accuracy previously achievable only for spectra from doubly charged precursors.

The higher sequencing accuracy in combined MS2/MS3 spectra increases the number of spectra for which one can recover long subsequences (i.e. tags) of at least six consecutive amino acids. The availability of MS3 ameliorates but does not completely eliminate the difficulties in de novo sequencing of complete peptides caused by missing fragment masses and high-intensity noise peaks. Nevertheless, one is often able to confidently recover tags that may uniquely identify the peptide using a simple (and very efficient) text-based database search. Even when these long tags happen to match multiple locations in larger databases, the number of possibilities tends to be very small and can be quickly resolved by matching the N/C-terminal masses (tags are usually recovered from the middle of the spectrum) and additional peaks not included in the tag. These factors are especially relevant when de novo sequencing is followed by homology-tolerant searches such as those enabled by MS-Alignment (Tout et al., 2005). In these cases, the tag-based efficiency gains (Bandeira et al., 2007h; Tanner et al., 2005) should combine with the improved ion statistics shown in Table 1d to simultaneously deliver faster and more accurate results.

4 CONCLUSION

High-throughput studies of unknown proteins such as recombinant (Pham et al., 2006) or venom (Birrell et al., 2007) proteins remain burdened by low-throughput experimental procedures such as Edman sequencing and the limited accuracy of de novo sequencing of individual tandem mass spectra. While related approaches increase the sequencing accuracy by combining multiple MS2 spectra from overlapping peptides (Bandeira et al., 2004; Bandeira et al., 2007a; Shevchenko et al., 1997), the experimental setup described here requires fewer sample handling steps (no need to digest with multiple enzymes) and should thus be applicable to smaller amounts of substrate.

We have demonstrated the utility of MS3 spectra for de novo peptide sequencing and compared the resulting gains with those obtainable from overlapping MS2 spectra. Although the ion statistics in individual MS3 spectra are usually too weak to allow reliable identification via database searching, the combination of multiple MS3 spectra with the parent MS2 spectrum results in higher signal-to-noise ratios and much improved separability of b/y-ions. By combining the corroborating fragmentation in multiple spectra our approach leads to increasing accuracy with higher numbers of usable MS3 spectra and even achieves almost error-free sequencing as soon as four usable MS3 spectra are available. The algorithmic incorporation of accurate parent masses was also shown to contribute to higher sequencing accuracy.

When applied to sets of overlapping MS2 spectra, the approach described here also resulted in sequencing accuracy improvements and practically erased the distinction between de novo sequencing of short and long peptides. These gains were even more pronounced when 5 overlapping MS2 spectra were available. An in-depth analysis of MS3 spectra (in the blue line, shewanella dataset) was due to inaccurate parent masses and was thus surpassed when accurate parent masses were available. Also, even with perfect sequencing accuracy, these spectra sometimes lacked tags of length six because of missing peaks, even though tags of length five were readily available.

---

6We note that the lower average sequencing accuracy with 5 overlapping MS2 spectra (in the blue line, shewanella dataset) was due to inaccurate parent masses and was thus surpassed when accurate parent masses were available. Also, even with perfect sequencing accuracy, these spectra sometimes lacked tags of length six because of missing peaks, even though tags of length five were readily available.

---

Multi-spectra peptide sequencing

aligned MS2/MS3 spectra (Fig. 2) allow one to separate between these ion types. Separating b/y-ions alleviates the uncertainties in peak ion-type assignments and thus reduces the probability of high-scoring incorrect peptide identifications. Furthermore, the parent masses of the MS2 spectra create a strong bias toward the corresponding b-ion masses. Since the overwhelming majority of all usable MS3 spectra come from true fragment masses, the resulting set of essentially mandatory b-/y-ions significantly reinforces the score of the correct peptide, while severely reducing the probability of an incorrect high-scoring peptide match (see Fig. S1 Supplementary Materials). This feature is very important for de novo-based approaches to peptide identifications (like RAId-DbS, Alves and Yu, 2005). The main bottleneck of such approaches is a large set of full-length suboptimal peptides that need to be generated to ensure that the correct peptide is searched against a database. As shown in Figure S1, multistage mass spectrometry has a potential to reduce the number of such peptides by several orders of magnitude. The gains obtained from merging MS3 spectra with the corresponding parent MS2 spectra are summarized in Table 1c.

Our analysis revealed that scored MS3 spectra sometimes have b/y-ions absent in MS2 spectra. Such additional fragmentation is especially important in a de novo sequencing context because one implicitly searches for the best peptide match over the space of all possible peptides of the observed parent mass. The combined contributions of the factors described above results in a strong bias toward the correct peptide sequence reflected in a significant increase in the average percentage of correctly predicted amino acids (from 85.7 to 90.7%, Fig. 4). Moreover, we note that this increase in sequencing accuracy is not achieved at the cost of making less predictions—sequencing MS2/MS3 spectra resulted in 4922 amino acid predictions versus 4772 for MS2 spectra only. As expected, having more usable MS3 spectra results in increasing de novo sequencing accuracy, generating almost-error-free sequences as soon as four usable MS3 spectra are available.

As illustrated by the blue lines in Figure 4, the gain in de novo sequencing accuracy is considerably higher when combining multiple MS2 spectra from overlapping peptides (shewanella dataset). The larger gains seem to be due to two distinct factors—(i) the lower average quality of MS2 spectra not yielding MS3 spectra (as compared to yeast MS2 spectra) and (ii) the better ion statistics of the MS2 spectra from overlapping peptides (as compared to yeast MS3 spectra). These two factors are clearly illustrated in Figure 4 by the differences between the blue/green dashed (case i) and full (case ii) lines. Nevertheless, increasing numbers of overlapping spectra also resulted in the same trend of better sequencing accuracy. In addition to these blue/green cases where the peptides from the yeast and shewanella datasets were most similar, we also observed significant improvements in the traditionally difficult de novo sequencing of long peptides (13+ amino acids). Since these peptides tend to generate less singly charged b/y-ion fragments in the MS2 spectrum (especially for triply charged precursors), these cases gain the most from the additional fragmentation in the additional spectra from overlapping peptides. Thus, the sequencing accuracy of long peptides becomes essentially indistinguishable from that of short peptides and doubly charged precursors. Also, while the sequencing accuracy on MS2 spectra from triply charged precursors still remains slightly inferior, it roughly approaches (or exceeds, when accurate parent masses are available) the sequencing accuracy previously achievable only for spectra from doubly charged precursors.
Fig. 4. Improvements in de novo peptide sequencing from overlapping spectra and accurate parent masses. (A) Changes in amino acid prediction accuracy, defined as percentage of correct amino acids out of all amino acid predictions; (B) Changes in number of spectra yielding a correct tag of six amino acids or longer (see Bandeira et al., 2007b, for a discussion of how these long tags could suffice for peptide identification). Four different scenarios were found to result in significantly different sequencing results (shown in four different colors), with the yeast dataset mostly resembling the blue subset of the shewanella dataset (charge two precursors with average peptide length ≈11.5). Values for single spectra are marked with triangle and the corresponding de novo sequencing gains (from additional overlapping spectra) are marked with a diamond of the same color and in the same column; values from the same peptide subset are connected by either dashed (single spectra) or full (multiple spectra) lines. The plots for the single peptides are not perfectly horizontal since there are small variations in the spectral quality of subsets of MS^2 spectra with i dependent MS^3 spectra (for 1 ≤ i ≤ 5).
The authors would like to thank Ari Frank and Sangtae Kim for insightful discussions.

Conflict of Interest: none declared.

REFERENCES

Acharya,R and Mann,M. (2003) Mass spectrometry-based proteomics. Nature, 422, 198–207.

Alves,G and Yu,Y.K. (2005) Robust accurate identification of peptides (raid): deciphering ms2 data using a structured library search with de novo based statistics.

Biointformatics, 21, 3726–3732.

Bafna,V and Edwards,N. (2003) On de-novo interpretation of tandem mass spectra for peptide identification. In Proceedings of the seventh annual international conference on Computational molecular biology, pp. 9–18.

Bandeira,N et al. (2004) Shotgun protein sequencing by tandem mass spectra assembly. Anal. Chem, 76, 7221–7233.

Bandeira,N et al. (2005) A shotgun protein sequencing: assembly of tandem mass spectra from mixtures of modified peptides. Mol. Cell Proteomics, 4, 1123–1134.

Bandeira,N et al. (2007b) Protein identification via spectral network analysis. Proc. Natl Acad. Sci. USA, 104, 6140–6145.

Barbie,C. (1990) Fast algorithm for peptide sequencing by mass spectrometry. BiomedICAL and Environmental Mass Spectrometry, 19, 363–368.

Birrell,G.W. et al. (1990) Fast algorithm for peptide sequencing by mass spectroscopy. J. Comput. Biol, 6, 325–337.

Busch,N et al. (1999) De novo peptide sequencing via tandem mass spectrometry. J. Comput. Biol, 6, 327–342.

Durbin,R et al. (1999) Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press.

Edwards,N et al. (2003) pepNovo: de novo peptide sequencing via probabilistic network modeling. Anal. Chem, 77, 964–973.

Frank,A.M and Pevzner,P.A. (2005) PepNovo: de novo peptide sequencing via probabilistic network modeling. Anal. Chem, 77, 964–973.

Fernández-de Cossío,J.

Edwards,N.J. (2007) Novel peptide identification from tandem mass spectra using ests and sequence database compression. Mol. Syst. Biol, 3, 102–102.

Fernández-de Cossío,J. et al. (1995) A computer program to aid the sequencing of peptides in collision-activated decomposition experiments. Comput. Appl. Biosci., 11, 427–434.

Fischer,R et al. (2005) Novohmm: a hidden markov model for de novo peptide sequencing. Anal. Chem, 77, 7265–7273.

Frank,A.M and Pevzner,P.A. (2005) PepNovo: de novo peptide sequencing via probabilistic network modeling. Anal. Chem, 77, 964–973.

Gupta,N et al. (2007) Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. Genome Res, 17, 1362–1377.

Haarum,J.S. (2006) Recombinant polyclonal antibodies: the next generation of antibody therapeutics? Drug. Discov. Today, 11, 655–660.

Kalkum,M et al. (2003) Detection of secreted peptides by using hypothesis-driven multistage mass spectrometry. Proc. Natl Acad. Sci. USA, 100, 2795–2800.

Lin,T and Gish,G.L. (1999) C-terminal peptide sequencing via multistage mass spectrometry. Anal. Chem, 70, 5162–5165.

McBride,M et al. (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid. Commun. Mass. Spectrom, 17, 2337–2342.

McIntyre,N (2005) Immunoglobulin gene diversification. Annu. Rev. Genet, 39, 23–46.

McLafferty,F et al. (2007) A dynamic programming algorithm for de novo peptide sequencing via tandem mass spectrometry. Anal. Chem, 79, 4870–4876.

Olsen,J.V and Mann,M. (2004) Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. Proc. Natl Acad. Sci. USA, 101, 13417–13422.

Perkins,D.N et al. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis, 20, 3551–3567.

Pevzner,P.A. et al. (2006) Performance evaluation of existing de novo sequencing algorithms. J. Proteome Res, 5, 3014–3028.

Pham,N et al. (2006) De novo proteomic sequencing of a monoclonal antibody raised against r50 lipagd. Anal. Biochem, 352, 77–86.

Pimentel,A.M and De Lima,M.E. (2005) Small peptides, big world: biotechnological potential in neglected biactive peptides from arthropod venoms. J. Pept. Sci, 11, 670–676.

Shevchenko,A et al. (1997) Rapid ‘de novo’ peptide sequencing by a combination of nanoelectrospray, isotopic labeling and a quadrupole-time-of-flight mass spectrometer. Rapid. Commun. Mass. Spectrom, 11, 1015–1024.

Tanner,S et al. (2005) InProTcT: identification of posttranslationally modified peptides from tandem mass spectra. Anal. Chem, 77, 4626–4639.

Tanner,S et al. (2007) Improving gene annotation using peptide mass spectrometry. Genome Res, 17, 231–239.

Tsai,D et al. (2005) Identification of post-translational modifications by blind search of mass spectra. Nat. Biotechnol, 23, 1562–1567.

Ulintz,P.J. et al. (2008) Investigating ms2/ms1 matching statistics: A model for coupling consecutive stage mass spectrometry data for increased peptide identification confidence. Mol. Cell Proteomics, 7, 61–67.

Vembu,A. (1987) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory, 33, 260–269.

Yates,J.R. et al. (1995) Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. Anal. Chem, 67, 5302–5310.

Zhang,Z and McLafferty,F.J. (2000) De novo peptide sequencing by two-dimensional fragment correlation mass spectrometry. Anal. Chem, 72, 2337–2350.