Deconvolution and Database Search of Complex Tandem Mass Spectra of Intact Proteins

A COMBINATORIAL APPROACH*

Xiaowen Liu‡, Yuval Inbar‡, Pieter C. Dorrestein§, Colin Wynne¶, Nathan Edwards∥, Puneet Souda**, Julian P. Whitelegge**, Vineet Bafna‡, and Pavel A. Pevzner‡ ‡

Top-down proteomics studies intact proteins, enabling new opportunities for analyzing post-translational modifications. Because tandem mass spectra of intact proteins are very complex, spectral deconvolution (grouping peaks into isotopomer envelopes) is a key initial stage for their interpretation. In such spectra, isotopomer envelopes of different protein fragments span overlapping regions on the \( m/z \) axis and even share spectral peaks. This raises both pattern recognition and combinatorial challenges for spectral deconvolution. We present MS-Deconv, a combinatorial algorithm for spectral deconvolution. The algorithm first generates a large set of candidate isotopomer envelopes for a spectrum, then represents the spectrum as a graph, and finally selects its highest scoring subset of envelopes as a heaviest path in the graph. In contrast with other approaches, the algorithm scores sets of envelopes rather than individual envelopes. We demonstrate that MS-Deconv improves on Thrash and Xtract in the number of correctly recovered monoisotopic masses and speed. We applied MS-Deconv to a large set of top-down spectra from Yersinia rohdei (with a still unsequenced genome) and further matched them against the protein database of related and sequenced bacterium Yersinia enterocolitica. MS-Deconv is available at http://proteomics.ucsd.edu/Software.html. Molecular & Cellular Proteomics 9.12:2772–2782, 2010.

Top-down proteomics is a mass spectrometry-based approach for identification of proteins and their post-translational modifications (PTMs)\(^1\) (1–14). Unlike the “bottom-up” approach where proteins are first digested into peptides and then a peptide mixture is analyzed by mass spectrometry, the top-down approach analyzes intact proteins. Thus, it has advantages in detecting and localizing PTMs as well as identifying multiple protein species (e.g. proteolytically processed protein species). Despite its advantages, top-down proteomics presents many challenges. These include requirement of high sample quantity, sophisticated instrumentation, protein separation, and robust computational analysis tools. For this reason, top-down proteomics has rarely been used for analyzing complex mixtures (12–18), and it is typically used to study single purified proteins. However, this situation is quickly changing with recent top-down studies of complex protein mixtures (14, 19).

Because of the existence of natural isotopes, fragment ions of the same chemical formula and charge state are usually represented by a collection of spectral peaks in tandem mass spectra called an isotopomer envelope. The monoisotopic mass of a chemical formula is the sum of the masses of the atoms using the principal (most abundant) isotope for each element. Spectral deconvolution focuses on grouping spectral peaks into isotopomer envelopes. By doing so, the charge state and monoisotopic mass of each envelope are effectively determined. A complex multi-isotopic peak list in the \( m/z \) space is translated into a simple monoisotopic mass list that is easier to analyze.

Given the monoisotopic mass and charge state of a fragment ion, its theoretical isotopic distribution can be predicted by assuming the fragment ion has an average elemental composition with respect to its mass (20) or using its precise elemental composition if the protein is known. Exploiting this, many deconvolution methods use theoretical isotopic distributions to detect and evaluate candidate isotopomer envelopes, which is the envelope detection problem (Fig. 1). To evaluate the fit of a candidate envelope to its theoretical isotopic distribution, many metrics have been proposed (20–32).

The candidate envelopes often overlap and share peaks, leading to a combinatorial problem of selecting the list of envelopes that best explains the spectrum (Fig. 2). In contrast to the well studied envelope detection problem, the envelope selection problem remains poorly explored. Most deconvolution algorithms follow a simple greedy approach to selecting the set of envelopes where the highest scoring envelopes are
iteratively selected and removed from the spectrum. Although this approach often generates reasonable sets of envelopes for simple spectra, its performance deteriorates in cases of complex spectra.

In particular, the greedy approach performs well when the envelopes are distributed sparsely along the m/z axis. Large proteins have many fragments that appear in multiple charge states. The high number of envelopes/peaks and the...
small m/z spread of the fragments with high charge states result in narrow m/z regions with high peak density. In these peak-dense regions, envelopes may overlap and share peaks, and the greedy approach and even manual interpretation often fail to find the optimal combination of envelopes (supplemental Fig. 1).

Several methods have been proposed to explore the envelope selection problem. Mcilwain et al. (33) presented a dynamic programming algorithm for selecting a set of envelopes such that the m/z ranges of the envelopes do not overlap. This non-overlapping condition becomes too restrictive for complex spectra of intact proteins. Samuelsson et al. (34) proposed a method that follows a non-negative sparse regression scheme. Du and Angeletti (35) and Renard et al. (36) addressed the envelope selection problem as a statistical problem of variable selection and used LASSO to solve it.

Here, we present MS-Deconv, a combinatorial algorithm for spectral deconvolution. MS-Deconv (i) generates a large set of candidate envelopes, (ii) constructs an envelope graph encoding all envelopes and relationships between them, and (iii) finds a heaviest path in the envelope graph. Although the envelope graph of a complex spectrum is large (exceeding a million nodes in some cases), the heaviest path algorithm can efficiently find an optimal set of envelopes. MS-Deconv explicitly scores combinations of candidate envelopes rather than individual envelopes as in previous approaches.

We tested MS-Deconv on a data set of top-down spectra from known proteins and evaluated the monoisotopic masses recovered by MS-Deconv. A mass was classified as a true positive if it was matched to the monoisotopic mass of a theoretical fragment ion of the protein within a specific parts per million (ppm) tolerance. We compared the performance of MS-Deconv with the widely used Thrash (20) and Xtract (37) and demonstrated that, with a few exceptions, MS-Deconv recovers more true positive masses. For example, for the collisionally activated dissociation (CAD) spectrum of bacteriorhodopsin (BR) with charge 10, the percentage of true positive mass peaks with intensity less than the noise intensity level are removed. Each peak with intensity larger or equal to the noise intensity level is considered as a candidate base peak. The range of charge states is from 1 to a user-defined parameter, called the maximum charge state. In practice, the maximum charge state can be defined arbitrarily, or one can scan MS1 spectra to estimate the charge state of the precursor ion and use this charge state as the maximum charge state. In difference from Thrash, we do not attempt to accurately define the charge state for each set of peaks but rather consider all feasible charge states and let the dynamic programming algorithm described under “Envelope Selection” to select the envelopes with charge states that make sense. Shared peaks are not taken into account when the candidate envelopes are generated. We emphasize that in difference from most deconvolution algorithms (e.g. Thrash) MS-Deconv works on centroided rather than profile spectra, resulting in an aggressive and inclusive envelope generation.

For each base peak and charge state, a candidate envelope $E$ is generated as follows. We start by generating a theoretical isotopic distribution $D$ such that the m/z value of its most intense peak is the same as that of the base peak. The distribution is obtained using the Emass (40) software, which calculates the expected isotopic distribution of a fragment ion (like Horn et al. (20), we assume the fragment ion has an average elemental composition with respect to its mass).

| Step | Description |
|------|-------------|
| 1. | Generate candidate envelopes |
| 2. | Construct an envelope graph representing candidate envelopes |
| 3. | Find a heaviest path in the envelope graph |
| 4. | Convert the heaviest path into a set of selected envelopes |
| 5. | Extract a monoisotopic mass from each selected envelope |

**Fig. 3. Main steps of MS-Deconv.**

**EXPERIMENTAL PROCEDURES**

Spectral Deconvolution as a Combinatorial Problem—Most existing methods address spectral deconvolution using a two-stage approach either explicitly or implicitly. The first stage is envelope detection: given an input spectrum, generate a set of candidate envelopes. This is followed by envelope selection: given a set of envelopes, find a subset of envelopes with the maximal score. Typically, the envelope selection problem is solved greedily by iteratively selecting the highest scoring envelope and further removing its peaks from the spectrum. In this study, we focus on a graph-theoretical approach to the envelope selection problem (illustrated in Fig. 3) that guarantees selecting a highest scoring set of envelopes.

Generating Candidate Envelopes—We briefly describe how candidate envelopes are generated prior to being scored and selected. We use the ReAdW (http://tools.proteomecenter.org/software.php) software for peak selection and centroiding and use an approach similar to Thrash (20) to estimate the noise intensity level in the centroided spectra by assuming that it is in the intensity bin with the largest number of peaks. After the noise intensity level is determined, all the peaks with intensity less than the noise intensity level are removed. Each peak with intensity larger or equal to the noise intensity level is considered as a candidate base peak. The range of charge states is from 1 to a user-defined parameter, called the maximum charge state. In practice, the maximum charge state can be defined arbitrarily, or one can scan MS1 spectra to estimate the charge state of the precursor ion and use this charge state as the maximum charge state. In difference from Thrash, we do not attempt to accurately define the charge state for each set of peaks but rather consider all feasible charge states and let the dynamic programming algorithm described under “Envelope Selection” to select the envelopes with charge states that make sense. Shared peaks are not taken into account when the candidate envelopes are generated. We emphasize that in difference from most deconvolution algorithms (e.g. Thrash) MS-Deconv works on centroided rather than profile spectra, resulting in an aggressive and inclusive envelope generation.

For each base peak and charge state, a candidate envelope $E$ is generated as follows. We start by generating a theoretical isotopic distribution $D$ such that the m/z value of its most intense peak is the same as that of the base peak. The distribution is obtained using the Emass (40) software, which calculates the expected isotopic distribution of a fragment ion (like Horn et al. (20), we assume the fragment ion has an average elemental composition with respect to its mass).
Next, the base peak and its neighboring spectral peaks are matched to the theoretical peaks in \( D \) by comparing their \( m/z \) values. The set of matched peaks within an error tolerance is denoted \( P \). Theoretical isotopic distribution \( D \) is scaled by comparing the total intensity of the base peak and its two neighboring peaks in \( P \) and their corresponding peaks in \( D \). We get a set of theoretical peaks \( T(\mathcal{E}) \) from the scaled distribution \( D \) by only keeping the most intense peaks with intensity larger or equal to the noise intensity level and the sum of their intensities just exceeding 85% of the total intensity. The set of peaks in \( P \) that are matched to the theoretical peaks in \( T(\mathcal{E}) \) is the candidate envelope \( \mathcal{E} \). \( T(\mathcal{E}) \) is called the pattern of \( \mathcal{E} \). Finally, each envelope is assigned to a 1-Da window, according to the location of its base peak. To reduce the number of candidate envelopes, at most, five highest scoring envelopes are selected in each 1-Da window. The filtering methods and scoring function are described in the supplemental material.

A candidate envelope is considered valid only if it satisfies some rather stringent requirements. Suppose the total number of peaks in a pattern is \( n \). A valid envelope of the pattern only allows at most one unmatched peak and requires it to have as least \( \max(3, \ n - 3) \) consecutive matched peaks. Using these constraints, most noise envelopes are removed from the candidate envelope list. Even though some noise envelopes remain in the candidate envelope list, most of them have lower scores than other envelopes assigned to the same 1-Da window. In most cases, they will be excluded from the output set of envelopes by the envelope selection algorithm described under “Envelope Selection” because it usually selects very few top scoring candidate envelopes in each 1-Da window. Thus, the filtering and envelope selection procedures work as de facto noise filters.

Envelope Selection—We assume that the envelope detection stage found a set of candidate envelopes \( \{E_1, E_2, \ldots, E_n\} \) along with a set of patterns \( \{T(E_1), T(E_2), \ldots, T(E_n)\} \). The score of an envelope \( \mathcal{E} \) is defined by \( s(\mathcal{E}) = \text{sim}(\mathcal{E}, T(\mathcal{E})) \) where \( \text{sim}(\mathcal{E}, T(\mathcal{E})) \) is a similarity score between \( \mathcal{E} \) and \( T(\mathcal{E}) \).

\[
s(\mathcal{E}) = \sum_{p \in \text{Peaks}(\mathcal{E})} \text{sim}(p, T(\mathcal{E})) \quad \text{(Eq. 1)}
\]

where \( \text{Peaks}(\mathcal{E}) \) is the collection of peaks in \( \mathcal{E} \) and \( \text{sim}(p, T(\mathcal{E})) \) is a scoring function between a peak \( p \) and a pattern \( T(\mathcal{E}) \). In the supplemental material, we describe the scoring function in detail.

We start by extending the score from a single envelope to a set of envelopes. Two envelopes are independent if they do not share peaks. For a set of envelopes \( A \), its score is defined as the sum of scores of all envelopes in the set if these envelopes are mutually independent and as \(-\infty\) otherwise.

\[
s(A) = \begin{cases} 
\sum_{\mathcal{E} \in A} s(\mathcal{E}), & \text{if all envelopes are mutually independent,} \\
-\infty, & \text{otherwise.}
\end{cases}
\quad \text{(Eq. 2)}
\]

We define a directed envelope graph based on the set of candidate envelopes (Fig. 4b). Let \( E_i \) be the set of all candidate envelopes containing an atom interval \( i \). For each subset \( X \subseteq E_i \), we generate a vertex \([i, X] \) if \( X \) is a set of mutually independent envelopes. Next, we add edges between the vertices. Vertices \([i, X] \) and \([i + 1, X'] \) in two neighboring atom intervals separated by a start/end \( a \) are connected by a directed edge from \([i, X] \) to \([i + 1, X'] \) if they satisfy one of the following three conditions. 1) \( X = X' \); 2) \( X \) is the start of \( E \), and \( X' = X \cup \{\mathcal{E}\} \); or 3) \( a \) is the end of \( E \), and \( X = X' \cup \{\mathcal{E}\} \). Finally, we assign weights to the vertices. The weight of a vertex \([i, X] \) is defined as \( s(\mathcal{E}) \) if the left end point of \( i \) is the start of an envelope \( \mathcal{E} \), and \( E \in X \). Otherwise, its weight is set to 0. All edge weights are set to 0.

The construction of the envelope graph implies that there is a one-to-one mapping between paths from the source (first atom interval) to the sink (last atom interval) in the graph and sets of mutually independent envelopes (see the supplemental material for the proof). The set of envelopes corresponding to a path is the union of the envelopes in all vertices of the path. For example, the heaviest path in Fig. 4c corresponds to the set of mutually independent envelopes \( \{E_1, E_3\} \). Moreover, the weight of a path equals the score of its corresponding envelope set. Thus, the envelope selection problem is reduced to the problem of finding a heaviest path in the envelope graph, which can be efficiently solved using a dynamic programming algorithm (41).

The number of vertices in the envelope graph is bounded by \((2n + 1)2^m\) where \( m \) is the maximal number of candidate envelopes containing an atom interval. Because each vertex has at most two incoming edges, the complexity of the algorithm is proportional to \( n2^m \). It turns out that \( m \) is small for many spectra in practice. Moreover, one can impose a constraint on \( m \) during the envelope detection stage.

The scenario depicted in Fig. 2b, where two envelopes share a peak, sometimes confounds the deconvolution of complex spectra. MS-Deconv has an option to use an intensity-split scoring model that allows peaks to be assigned to multiple envelopes. When a set of envelopes is evaluated, the intensity of a peak is virtually distributed between all envelopes that share the peak, according to the intensities of its corresponding peaks in the intensity-scaled theoretical isotopic distributions. If an observed envelope fits poorly to the theoretical distribution because it shares peaks with other envelopes, it might fit well to the theoretical distribution using the intensity-split model. In the supplemental material, we redefine the scoring function for the case when selected envelopes are allowed to share peaks.

We use the dynamic programming algorithm to select a subset of envelopes from the candidate envelope set. To extract the monoisotopic masses of the selected envelopes, we define a distance between a theoretical isotopic distribution and an experimental envelope. This function is used to address the notoriously difficult problem of correcting ±1-Da errors in the list of monoisotopic masses (see the supplemental material for details).

RESULTS AND DISCUSSION

Data Sets—We tested MS-Deconv, Thrash, and Xtract (Thermo Scientific) on a collection of six CAD spectra of two intact proteins: BR from *Halobacterium* (P02945) and apolipoprotein A-I (apoA-I) from pig that carries the “HV” to “QL” variation (P18648). BR contains 248 amino acids (after removing its N-terminal peptide “MLELPTAVEGVS” and C-terminal “Asp”), and its monoisotopic mass is 26,766.12 Da (with loss of an ammonia at its N terminus). ApoA-I contains 241 amino acids (after removing its N-terminal peptide “MIKAVVTLAVLFLTGSQARHFWQQ”), and its monoisotopic mass is 27,866.22 Da. Top-down mass spectrometry was performed.
using CAD (the same method as described (10, 11)). Three of the six spectra are from BR with charges 10, 11, and 16. The other three are from apoA-I with charges 23, 25, and 26. We further tested MS-Deconv in conjunction with a protein database search of top-down spectra from *Y. rohdei*. The spectral data set was acquired from a top-down LC-MS/MS experiment on an LTQ-Orbitrap (ThermoFisher). The experimental procedure is described in Wynne et al. (19). The precursor masses range from 5000 to 20,000 Da, and the charge states range from 3 to 15. We used ReAdW to convert Thermo raw files to mzXML data files. To improve the quality of MS/MS spectra, we merged similar MS/MS spectra presumed to correspond to the same protein. MS/MS spectra are merged if their precursor ions have the same charge state/monoisotopic mass and if they share most peaks. We ran MS-Deconv to generate a monoisotopic mass list for each spectrum and focused our attention on 331 spectra with at least 20 monoisotopic masses.

Comparison among MS-Deconv, Thrash, and Xtract—We compared MS-Deconv with Thrash (20) (run through ProSightPC). The output of MS-Deconv is a collection of envelopes (sorted by score) and their monoisotopic masses. Thrash outputs a list of monoisotopic masses that are not explicitly assigned to envelopes. To ensure a fair comparison between the two tools, we compared them based on the output monoisotopic mass lists only.

Thash does not report the score of each output monoisotopic mass and thus does not allow one to rank these masses. Although the absence of ranking is a deficiency of Thrash (such ranking is useful for protein identifications), we nevertheless made an attempt to compare MS-Deconv with Thrash by emulating such ranking using the minimal RL parameter in Thrash that reflects the fit between the theoretical and experimental isotopic distributions. Although it is not a perfect way to evaluate Thrash, varying the parameter RL value leads to generating varying numbers of monoisotopic masses. To ensure a comprehensive benchmarking, we ran Thrash with the default RL value 0.9 as well as several other values. By contrast, MS-Deconv can rank the output envelopes and report the same number of monoisotopic masses as Thrash.

To evaluate the MS-Deconv and Thrash results, we generate for each spectrum a theoretical monoisotopic mass list from the known protein sequence. Each list consists of the monoisotopic masses of b-ions and y-ions plus some masses with some offsets (detailed in supplemental Section 8 and Fig. 2). For each mass in the output mass list, we assume it is a true positive mass if its nearest theoretical mass is within a given ppm error tolerance. In the comparison, ppm thresholds 3, 5, and 10 are used.
Prior to benchmarking, we performed a linear recalibration of all spectra (detailed in supplemental Section 9 and Fig. 3). To ensure fairness, the same recalibration procedure was applied to both MS-Deconv and Thrash results. We then assessed the recalibrated mass lists using the same evaluation procedure for both deconvolution methods.

Table I describes the MS-Deconv and Thrash results for six spectra on three RL values and three values of ppm each (54 tests overall). In 44 of 54 tests, MS-Deconv improved on Thrash (20% increase in the number of true positive masses on average). Table I illustrates that MS-Deconv performs better than Thrash on complex spectra with many peaks where it gained 50% more true positive masses in some cases. Fig. 5 shows the number and percentage of true positive masses (true positive rate) in the output mass lists for the spectrum of BR with charge 10 (similar curves for other spectra are shown in supplemental Fig. 4). The percentage of true positive masses among the top 150 masses is above 70% for MS-Deconv and less than 50% for Thrash. One of the possible reasons why the false positive rate is so high (and increases even further with increasing the number of selected envelopes) may be explained by the stringent requirement for considering a mass as a true positive. Some masses may be qualified as false positives because they represent internal fragment ions, uncommon neutral losses, or multiple PTMs.

We also compared MS-Deconv with Xtract (37). Similar to Thrash, the output of Xtract is a list of monoisotopic masses. We ran Xtract with the minimum fit parameter set to 65, 75, and 85 and MS-Deconv reporting the same number of monoisotopic masses as Xtract. Thresholds 3, 5, and 10 ppm

| Sp. | Protein | Charge | No. peaks | RL value | No. output masses |
|-----|---------|--------|-----------|----------|------------------|
| 1   | BR      | 10     | 20,125    | 0.9999   | 150              |
|     |         |        |           | 0.9      | 272              |
|     |         |        |           | 0.5      | 342              |
|     |         |        |           | 0.5      | 734              |
| 2   | BR      | 11     | 19,916    | 0.9999   | 238              |
|     |         |        |           | 0.9      | 597              |
|     |         |        |           | 0.5      | 734              |
| 3   | BR      | 16     | 6,389     | 0.9999   | 41               |
|     |         |        |           | 0.9      | 106              |
|     |         |        |           | 0.5      | 159              |
| 4   | ApoA-I  | 23     | 8,141     | 0.9999   | 77               |
|     |         |        |           | 0.9      | 180              |
|     |         |        |           | 0.5      | 269              |
| 5   | ApoA-I  | 25     | 7,917     | 0.9999   | 55               |
|     |         |        |           | 0.9      | 132              |
|     |         |        |           | 0.5      | 223              |
| 6   | ApoA-I  | 26     | 5,605     | 0.9999   | 38               |
|     |         |        |           | 0.9      | 75               |
|     |         |        |           | 0.5      | 143              |
were used to determine true positive masses. The results of the comparison are reported in supplemental Table 1. In 53 of 54 tests, MS-Deconv improved on Xtract (∼29% increase in the number of true positive masses on average). However, it is somewhat unfair to compare MS-Deconv (or Thrash) with Xtract because Xtract combines close monoisotopic masses identified from different envelopes into a single mass. Thus, MS-Deconv and Thrash are compared based on the number of recovered envelopes, whereas Xtract is compared based on the number of recovered monoisotopic masses.

We also compared the running time of MS-Deconv, Xtract, and Thrash. The tools were tested on a PC with a 2.2-GHz CPU and 3.0-GB RAM. The running time (average over six spectra) for MS-Deconv, Xtract, and Thrash is 9, 36, and 302 s, respectively. The input of Xtract and Thrash is Thermo raw data files, and the input of MS-Deconv is mzXML files extracted from Thermo raw data files. The average time for converting the raw files to mzXML files is less than a second.

Searching Protein Databases with Top-down Spectra—We searched the protein database of Y. enterocolitica against the top-down spectral data set of Y. rohdei. Y. enterocolitica represents one of the most similar sequenced genomes in comparison with Y. rohdei. We used the protein database of Y. enterocolitica as a proxy for the protein database of Y. rohdei and compared all spectra from Y. rohdei against all proteins in Y. enterocolitica with the goal to detect proteins in Y. rohdei that represent mutated or modified versions of proteins in Y. enterocolitica.

We scanned the MS1 spectra and used the scoring function for candidate envelopes in MS-Deconv to determine the charge state and monoisotopic mass of each precursor ion. Then we ran MS-Deconv to extract a monoisotopic mass list from each of the MS/MS spectra. We focused our analysis on spectra with long mass lists (20 or more masses), leaving us with 331 mass lists. We also ran Thrash to process the same data set with its default setting.

Wynne et al. (19) identified 10 proteins in Y. rohdei using Thrash and ProSightPC to search Swiss-Prot proteins from Yersinia species for exactly conserved protein sequences (eight of 10 proteins were identified using a database search of Y. enterocolitica). Comparing Wynne et al. (19) and our results is complicated by the fact that similar or identical proteins from different Yersinia species can be matched to the same spectrum. Thus, we selected the proteome of a single species, Y. enterocolitica, as the proxy database and ignored two proteins identified by comparison with proteomes of other Yersinia species.

We used the spectral alignment algorithm (39, 42) to compare the mass lists reported by MS-Deconv and Thrash. This algorithm can identify proteins with multiple PTMs in a protein database using a monoisotopic mass list. The advantage of this approach (as compared with ProSightPC) is that it can capture unknown PTMs and mutations. Our spectral alignment algorithm for comparing a top-down spectrum against a protein is similar to the algorithm described in Frank et al. (39). We generate an extended mass list from the reported monoisotopic mass list. For each mass m in the list, we add two masses m and parent mass – m to the extended mass list (parent mass is predicted from MS1 spectra). We also generate a list of theoretical monoisotopic masses of N-terminal ions for each protein in the protein database. An alignment between an extended monoisotopic mass list and a theoretical mass list can be visualized as a path in a two-dimensional grid (39) consisting of diagonal segments connected by vertical and horizontal segments representing mass shifts. Each mass shift corresponds to a PTM or a mutation. Examples are shown in supplemental Fig. 5.) The number of matching mass pairs in the path (±1-Da errors are allowed) is defined as the score of the spectral alignment. We used a dynamic programming algorithm from Frank et al. (39) to find the optimal alignment between the extended and theoretical mass lists. The mass shifts at the N terminus and C terminus of the protein can be arbitrary (to account for truncated proteins) but are limited to ±300 Da for the mass shifts on internal residues. We ran spectral alignment with at most two mass shifts (PTMs or mutations), ±2-Da error tolerance for the parent mass, and 15-ppm fragment ion error tolerance. For each monoisotopic mass list, we reported the protein sequence with the largest number of matched masses (see supplemental Tables 2–17 and Figs. 6 and 7).

We used the spectral alignment algorithm to search the mass lists extracted from the Y. rohdei spectral data set (331 spectra) by MS-Deconv against the protein database of Y. enterocolitica. This search identified eight proteins (Table II) containing amino acid substitutions and methylation PTMs not reported in Wynne et al. (19), which used ProSightPC to search for exactly conserved protein sequences in related organisms.

For each of the eight spectra identified in Wynne et al. (19), we removed the low ranking masses from the mass list of MS-Deconv so that the two mass lists reported by MS-Deconv and Thrash have the same number of masses. We searched the eight pairs of mass lists from MS-Deconv and Thrash against the protein database of Y. enterocolitica using the spectral alignment algorithm and identified the same eight proteins as in Wynne et al. (19). Although the NME modifications observed here can be readily precomputed by ProSightPC, we point out that the spectral alignment algorithm discovers all modifications in the blind mode. The numbers of matched fragments are reported in Table III under the “Alignment” heading.

We further compared the same eight pairs of mass lists (from MS-Deconv and Thrash) by running ProSightPC against the protein database of Y. enterocolitica. The precursor ion mass tolerance was set to 2.5 Da, and the fragment ion mass tolerance was set to 15 ppm. The numbers of matched fragments are reported in Table III under the “ProSightPC” heading. The spectral alignment algorithm often reveals more
Deconvolution of Complex Spectra

**TABLE II**

Combining MS-Deconv with spectral alignment to identify new proteins

MS-Deconv is used to deconvolute spectra from *Y. rohdei* that are further searched against *Y. enterocolitica* proteome with the spectral alignment algorithm. Eight proteins are identified that are not reported in Wynne et al. (19). The numbers of matched fragments and masses, mass shifts, and possible PTMs and mutations are reported. In the last column, we list a tentative assignment of mutations and/or PTMs. Sp., spectrum.

| Sp. | Monoisotopic mass (Da) | Charge | Protein | No. reported masses | No. matching masses | No. matching fragments | Mass shifts | Possible PTMs or mutations |
|-----|------------------------|--------|---------|---------------------|-------------------|-----------------------|------------|-----------------------------|
| 1   | 11,206.29              | 11     | A1JS21  | 109                 | 53                | 40                    | -161, +30  | NME and Ser → Gly, Ala → Thr |
| 2   | 9,567.14               | 12     | A1JS24  | 85                  | 42                | 29                    | -143       | NME and Val → Ser           |
| 3   | 9,629.17               | 11     | A1JU4   | 81                  | 37                | 34                    | +12, +16   | Ser → Val, Ala → Ser        |
| 4   | 9,033.90               | 12     | A1JK22  | 241                 | 51                | 36                    | -86        | Lys → Gly and Glu → Asn     |
| 5   | 9,245.97               | 13     | A1JNM8  | 216                 | 66                | 49                    | +14        | Methylation                 |
| 6   | 10,276.70              | 13     | A1JS33  | 85                  | 38                | 26                    | -131, -16  | NME, Ser → Ala              |
| 7   | 10,905.79              | 9      | A1JRC9  | 98                  | 63                | 49                    | -16        | Glu → Ile                   |
| 8   | 6,236.55               | 10     | A1JHR4  | 97                  | 36                | 25                    | -117       | NME and methylation         |

**TABLE III**

Comparison between MS-Deconv and Thrash in conjunction with ProSightPC and spectral alignment algorithm

The monoisotopic mass lists reported by MS-Deconv and Thrash (with the same numbers of masses) from eight spectra identified in Wynne et al. (19) are searched against a protein database of *Y. enterocolitica* using ProSightPC and the spectral alignment algorithm. The numbers of matched fragments in the mass lists reported by MS-Deconv and Thrash are compared. Sp., spectrum.

| Sp. | Monoisotopic mass (Da) | Charge | Protein | No. matching fragments | No. matching masses | Mass shift | Possible PTM |
|-----|------------------------|--------|---------|------------------------|-------------------|------------|--------------|
| 1   | 8,857.76               | 13     | A1JH2R  | 65                     | 13                | 16         | NME         |
| 2   | 6,041.17               | 8      | A1JN60  | 57                     | 28                | 27         | NME         |
| 3   | 8,363.65               | 11     | A1JQX1  | 32                     | 10                | 10         | NME         |
| 4   | 6,410.58               | 8      | A1JS10  | 93                     | 20                | 22         | NME         |
| 5   | 7,256.94               | 9      | A1JJS2  | 92                     | 26                | 24         | NME         |
| 6   | 6,848.67               | 8      | A1J11   | 31                     | 12                | 9          | NME         |
| 7   | 12,148.77              | 11     | A1JS31  | 31                     | 5                  | 8          | NME         |
| 8   | 14,988.35              | 13     | A1JIS8  | 31                     | 14                | 8          | NME         |

* A1JQX1 is not a complete identification. It is included because it was reported in Wynne et al. (19).

matched fragments than ProSightPC because it accounts for ±1-Da errors.2

MS-Deconv and Thrash can be evaluated by the number of matched fragments reported by database search tools like ProSightPC and spectral alignment. Although MS-Deconv and Thrash result in similar performance with ProSightPC, MS-Deconv results in more matched fragments than Thrash in conjunction with the spectral alignment. This indicates that MS-Deconv and Thrash report similar numbers of matched masses (without accounting for ±1-Da errors), but MS-Deconv also reports some (valuable) matched masses with ±1 Da. Because ProSightPC does not utilize these masses, the combination of MS-Deconv and spectral alignment becomes more powerful in database searches.

Predicting exact monoisotopic masses is a notoriously difficult problem because the theoretical envelopes for masses differing by a single dalton are very similar. As a result, both Thrash and MS-Deconv sometimes make ±1-Da errors while predicting monoisotopic masses. Common modifications like oxidation (with +16-Da offset) may appear as unusual modifications with +15- or +17-Da offset in the subsequent protein database searches. As discussed in Frank et al. (39), one has to be careful in interpreting the modifications returned by top-down searches. We manually inspected the reported alignments to select the mass shifts that are consistent with mutations or common PTMs (see the supplemental material for details).

The mass shifts in eight newly identified proteins from *Y. rohdei* are caused by mutations or PTMs that are not accounted for in ProSightPC. This analysis illustrates that MS-Deconv in conjunction with the spectral alignment has an ability to identify proteins with unknown mutations and PTMs. The detailed information about these additional proteins is reported in the supplemental material.

2 In some cases, the results of the spectral alignment algorithm have fewer matched fragments than those of ProSightPC. The reason is that the spectral alignment algorithm requires a relatively high accuracy of reported masses to find a good alignment. If the accuracy of the reported masses is low, the number of matched fragments identified by the spectral alignment algorithm will decrease.
Visualization of Envelopes, Protein Annotation, and Peptide Sequence Tags—MS-Deconv outputs images of envelopes (shown in supplemental Fig. 8) as well as images of each 10 m/z interval of the annotated spectrum (Fig. 6a) to enable manual inspections of the results.

The monoisotopic mass list reported by MS-Deconv can be used for protein sequence annotation or peptide sequence tag prediction dependent on whether the protein sequence is known or not. If the protein sequence is known, we can map the masses to its theoretical fragment ions of the sequence. The annotated BR protein sequence with the list of 342 masses recovered from the spectrum of BR with charge 10 is shown in Fig. 6b. Of the 247 breakage points, 45 points are supported by at least one reported mass of N-terminal ion, 36 points are supported by at least one reported mass of C-terminal ion, and 27 points are supported by reported masses of both N-terminal and C-terminal ions. The list with the same number of masses reported by Thrash covers only 36 points for N-terminal ions, 32 points for C-terminal ions, and 23 points for both N-terminal and C-terminal ions.

If the protein sequence is unknown, we generate peptide sequence tags (see supplemental Table 18 for details). For the spectra of BR with charges 10 and 11, the tags found are similar to the correct peptide sequence tags. Most errors are due to the 1-dalton mass shift introduced in the derivation of monoisotopic masses from isotopomer envelopes.

Acknowledgments—Data were collected at the University of California Los Angeles. We thank Professor Joseph Loo for access to the LTQ-FT instrument, purchased with National Institutes of Health Grant S10 RR023045.
of < 60-kDa proteins from Methanosarcina acetivorans using quadrupole FRMS with automated octopole collisionally activated dissociation. Mol. Cell. Proteomics 5, 14–25

16. Sharma, S., Simpson, D. C., Tolić, N., Jaity, N., Mayumaparthi, A. M., Smith, R. D., and Pasa-Tolić, L. (2007) Proteomic profiling of intact proteins using WAX-RPLC 2-d separations and FTICR mass spectrometry. J. Proteome Res. 6, 602–610

17. Wynne, C., Fenselau, C., Demirev, P. A., and Edwards, N. (2009) Top-down identification of protein biomarkers in bacteria with unsequenced genomes. Anal. Chem. 81, 9633–9642

18. Horn, D. M., Zubarev, R. A., and McLafferty, F. W. (2000) Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. J. Am. Soc. Mass Spectrom. 11, 330–332

19. Szymura, J. A., and Lamkiewicz, J. (2003) Band composition analysis: a new procedure for deconvolution of the mass spectra of organometallic compounds. J. Mass Spectrom. 38, 817–822

20. Wehofsky, M., Hoffmann, R., Hubert, M., and Spengler, B. (2001) Isotopic deconvolution of matrix-assisted laser desorption/ionization mass spectra for substance-class specific analysis of complex samples. Eur. J. Mass Spectrom. 7, 39–46

21. Gras, R., Müller, M., Gasteiger, E., Gay, S., Binz, P. A., Bienvenut, W., Hoogland, C., Sanchez, J. C., Bairoch, A., Hochstrasser, D. F., and Appel, R. D. (1999) Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. Electrophoresis 20, 3535–3540

22. Szymura, J. A., and Lamkiewicz, J. (2003) Band composition analysis: a new procedure for deconvolution of the mass spectra of organometallic compounds. J. Mass Spectrom. 38, 817–822

23. Mason, C. J., Therneau, T. M., Eckel-Passow, J. E., Johnson, K. L., Oberg, A. D., Loo, J. A., Wilkins, M. R. (2008) Automated Poisson peak harvesting for high throughput protein identification. Electrophoresis 21, 2243–2251

24. Chen, L., Szé, S. K., and Yang, H. (2006) Automated intensity descent algorithm for interpretation of complex high-resolution mass spectra. Mol. Cell. Proteomics 9, 3621–3631

25. Wang, W., Zhou, H., Lin, H., Roy, S., Shaler, T. A., Hill, L. R., Norton, S., Kumar, P., Anderle, M., and Becker, C. H. (2003) Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. Anal. Chem. 75, 4818–4826

26. Park, K., Yoon, J. Y., Lee, S., Paek, E., Park, H., Jung, H. J., and Lee, S. W. (2008) Isotopic peak intensity ratio based algorithm for determination of isotopic clusters and monoisotopic masses of polypeptides from high-resolution mass spectrometric data. Anal. Chem. 80, 7294–7303

27. McLwaine, S., Page, D., Hufton, E. L., and Sussman, M. R. (2007) Using dynamic programming to create isotopic distribution maps from mass spectra. Bioinformatics 23, 3328–3336

28. Samuelsson, J., Dalevi, D., LeDuc, R. D., and Glovaczk, K. J., Kim, Y. B., Viswanathan, V.,
Spaulding, I. T., Early, B. P., Bluhm, E. J., Babai, S., and Kelleher, N. L. (2007) ProSight PTM 2.0: improved protein identification and characterization for top-down mass spectrometry. *Nucleic Acids Res.* **35**, W701–W706

39. Frank, A. M., Pesavento, J. J., Mizzen, C. A., Kelleher, N. L., and Pevzner, P. A. (2008) Interpreting top-down mass spectra using spectral alignment. *Anal. Chem.* **80**, 2499–2505

40. Rockwood, A. L., and Haimi, P. (2006) Efficient calculation of accurate masses of isotopic peaks. *J. Am. Soc. Mass Spectrom.* **17**, 415–419

41. Jones, N. C., and Pevzner, P. A. (2000) *An Introduction to Bioinformatics Algorithms*, MIT Press, Cambridge, MA

42. Tsur, D., Tanner, S., Zandi, E., Bafna, V., and Pevzner, P. A. (2005) Identification of post-translational modifications by blind search of mass spectra. *Nat. Biotechnol.* **23**, 1562–1567