Peptide Identification by Database Search of Mixture Tandem Mass Spectra*

Jian Wang‡, Philip E. Bourne§, and Nuno Bandeira§¶**

In high-throughput proteomics the development of computational methods and novel experimental strategies often rely on each other. In certain areas, mass spectrometry methods for data acquisition are ahead of computational methods to interpret the resulting tandem mass spectra. Particularly, although there are numerous situations in which a mixture tandem mass spectrum can contain fragment ions from two or more peptides, nearly all database search tools still make the assumption that each tandem mass spectrum comes from one peptide. Common examples include mixture spectra from co-eluting peptides in complex samples, spectra generated from data-independent acquisition methods, and spectra from peptides with complex post-translational modifications. We propose a new database search tool (MixDB) that is able to identify mixture tandem mass spectra from more than one peptide. We show that peptides can be reliably identified with up to 95% accuracy from mixture spectra while considering only a 0.01% of all possible peptide pairs (four orders of magnitude speedup). Comparison with current database search methods indicates that our approach has better or comparable sensitivity and precision at identifying single-peptide spectra while simultaneously being able to identify 38% more peptides from mixture spectra at significantly higher precision. *Molecular & Cellular Proteomics* 10: 10.1074/mcp.M111.010017, 1–11, 2011.

Over the past several years there have been substantial advances in the sensitivity of protein identification approaches thanks to technological developments in chromatography (1) and tandem mass spectrometry (2, 3). In shotgun proteomics, researchers can routinely identify thousands of proteins from complex biological samples in a single experiment (4–6). But, despite this rapid progress, there are still challenging issues that remain unsolved (7, 8). One such challenge is that in any high throughput tandem MS (MS/MS) experiment only a fraction of MS/MS spectra can be identified by current computational methods. Although there are many factors contributing to this low spectrum identification rate, recent studies suggest that one reason is the occurrence of co-eluting peptides. Several analyses showed that in any complex mixture (e.g. whole-cell lysate), peptides with similar masses and chromatographic properties can occur quite frequently and give rise to mixture spectra generated from more than one peptide (9–12). Mixture spectra can confuse current computational methods because most mainstream approaches make the assumption that each MS/MS spectrum comes from only one peptide. Houel et al. estimated that identification rates for mixture spectra can be half of what is expected for single-peptide spectra (12). Mixture spectra have also been explored by alternative data acquisition approaches to overcome the limitations of current mass spectrometry scan rates (13–16).

In order to alleviate the algorithmic bottleneck in the identification of mixture spectra, Zhang et al. described the first database search tool for mixture spectra, ProbiDtrees (17) and showed that it is possible to identify co-eluting peptides from single MS/MS spectra. However, in that study not all peptides could be confidently identified and in most cases ProbiDtrees only identified the most prominent peptide in each spectrum. Recently, we also proposed a spectral library search tool, M-SPLIT, and showed that it is able to reliably and efficiently identify peptides from mixture spectra from co-eluting peptides in a yeast cell lysate (18). Despite the rapid growth of publicly available spectral libraries (19, 20), methods based on spectral matching suffer from the limitation that peptides cannot be identified if they have not been observed before; thus database search methods are still the mainstream approach for peptide identification. Some database search tools approach the mixture spectra identification problem by reporting spectra with more than one significant single-peptide match and do not explicitly attempt to model the occurrence of fragment ions from two peptides in the same spectrum. False discovery rates (FDR) are also left unadjusted (17, 21, 22) and

---

*The abbreviations used are: MS/MS, tandem mass spectrometry; PTM, post-translational modification; M-SPLIT, mixture-spectrum partitioning using libraries of identified tandem mass spectra; PSM, peptide spectrum match; PPSM, peptide/peptide spectrum match; FDR, false discovery rate; SVM, support vector machine.

---

Want to cite this article? Please look on the last page for the proper citation format.
may result in higher than expected FDR for second IDs (e.g. when co-eluting peptides share a substantial number of fragment masses). Different from previous approaches, our new database search tool, MixDB, uses a scoring model specifically designed for matching spectra against pairs of peptides and determines separate FDRs for identification of single-peptide spectra and mixture spectra. Applying our method to a yeast lysate data set, we show that using efficient filtration techniques and a rigorous probabilistic scoring model, peptides can be reliably identified from mixture spectra while only considering a small fraction of all possible peptide pairs.

**EXPERIMENTAL PROCEDURES**

**Problem Formulation**—A mixture spectrum is defined as an MS/MS spectrum from two different peptides. Analogous to the identification of single-peptide MS/MS spectra by comparison against all peptides in a database of known protein sequences, our goal is to identify mixture spectra by comparison against all possible pairs of peptides in a given protein sequence database. More formally, we represent an MS/MS spectrum as a real-value vector \( \mathbf{S} = s_1, s_2, \ldots, s_n \) where each element corresponds to the total peak intensity at a particular mass bin (bin size depends on instrument resolution). A mixture spectrum \( \mathbf{M} \) is modeled as \( \mathbf{M} = \mathbf{A} + \alpha \mathbf{B} \), where \( \mathbf{A} \) and \( \mathbf{B} \) are MS/MS spectra from two different peptides and \( \alpha \), the mixture coefficient, indicates their relative abundance. Without loss of generality, we assume that \( \mathbf{A} \) and \( \mathbf{B} \) are scaled to the same magnitude and that \( 0 \leq \alpha \leq 1 \) (i.e. \( \mathbf{A} \) always corresponds to the higher abundance peptide). We define a peptide sequence as a vector \( \mathbf{P} = p_1, p_2, \ldots, p_n \), where \( p_i \) is nonzero if there is at least one theoretical ion mass in the corresponding mass bin. A database \( \mathcal{D} \) is simply a set of peptides \( \mathcal{D} = \{ \mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_m \} \). We can now formulate the following computational problem:

**Mixture Spectrum Identification Problem (MSIP)**—

**Input**—A putative mixture spectrum \( \mathbf{M} \) and a sequence database \( \mathcal{D} \).

**Output**—A set of peptide pairs \( \{ \mathbf{P}_i, \mathbf{P}_j \} \in \mathcal{D} \), maximizing \( \text{PPSM}(\mathbf{M}, \mathbf{P}_i, \mathbf{P}_j) \) where \( \text{PPSM} \) is a given Peptide/Peptide Spectrum Match scoring function that describes how well a pair of peptides \( \mathbf{P}_i, \mathbf{P}_j \) matches the spectrum \( \mathbf{M} \).

Because there is currently no publicly available data with validated identifications of mixture MS/MS spectra, we created a data set of simulated mixture spectra to develop and benchmark our approach using the procedure described in (18). Briefly, representative MS/MS spectra of peptides with precursor charge 2 or 3 are selected from the NIST yeast spectral library\(^2\) (23). All spectra are first scaled to the same total intensity by dividing each peak intensity by the sum of all peak intensities in the spectrum. Because in a mixture the two peptides will most likely be present at different abundances, mixture spectra were created by randomly selecting two spectra from the library and linearly combining them using \( \mathbf{M} = \mathbf{A} + \alpha \mathbf{B} \) with predefined mixture coefficients \( 0 \leq \alpha \leq 1 \). Because co-eluting peptides have similar precursor mass, we further required that spectrum pairs have precursor mass difference less than 3 Th. Noise was assumed to be additive in mixture spectra and thus modeled by not removing unexplained/noise peaks from single-peptide spectra used to simulate mixture spectra. To simulate signal suppression, the final spectra were filtered by keeping only the eight most intense peaks in every 50 Da window.

\(^2\) There were also spectra with charge 1 and 4 in the library, but because the number of spectra was not enough to train a statistical model for the scoring function those were not considered in this study.

**Filtration Strategy**—Although the MSIP formulation is simple, the large number of possible peptide candidates in a sequence database \((10^4–10^5)\) peptides in the yeast database after precursor mass filtering with 3 Da tolerance) makes searching all possible pairs of peptides a prohibitive approach: the resulting \( \sim 10^{10} \) comparisons per query spectrum would mean that a typical data set of 15,000 spectra would take \( \sim 50 \) days to process (based on average InsPeCT runtimes, previously shown to be \( \sim 100 \) times faster than SEQUEST (24)), even without considering the additional computational burden of scoring spectra against pairs of peptides. The explosion in the number of candidate matches per spectrum also dramatically increases the chances for false-positive matches. We propose a filtration strategy to avoid this quadratic penalty of searching all pairs by first scanning through the database to discard peptide candidates that are not likely to result in good mixture matches and then pair up the remaining candidates to find the best scoring peptide pairs.

Core to any database search method is a peptide spectrum match (PSM) scoring function to describe the quality of a match between a candidate peptide and a given spectrum. Conceptually, this scoring function typically gives some reward for matches between observed peaks in the spectrum and theoretical fragment masses from the candidate peptides and imposes penalties for unexplained spectrum peaks or missing theoretical ion masses from the candidate peptide. When scanning the database to select promising peptide candidates, one needs to consider that a mixture spectrum \( \mathbf{M} \) derived from peptides \( \mathbf{A} \) and \( \mathbf{B} \) may have a poor PSM score to either peptide alone; that is, the presence of \( \mathbf{B} \) in the mixture spectrum typically results in many unmatched peaks between \( \mathbf{M} \) and \( \mathbf{A} \). We address this issue with the concept of *projected-spectrum* (18), in which the match between \( \mathbf{M} \) and \( \mathbf{A} \) only considers peaks in \( \mathbf{M} \) if there is a theoretical ion from \( \mathbf{A} \) with the same mass (within peak mass tolerance). More precisely, for a spectrum \( \mathbf{M} \) and a peptide \( \mathbf{A} \), let \( \mathbf{M}_{pA}[i] \) denote the \( p \)-th bin in spectrum \( \mathbf{M} \), the projection of \( \mathbf{M} \) onto \( \mathbf{A} \) \( \mathbf{M}_{pA}[i] \) is defined as:

\[
M_{pA}[i] = \begin{cases} 
M[i] & \text{if } A[i] > 0 \\
0 & \text{otherwise}
\end{cases}
\]

The intuition behind this operation is that because the peptide \( \mathbf{A} = a_1, a_2, \ldots, a_n \) contains all theoretical ion masses generated from this peptide, the projection essentially separates the peaks that are possibly generated from peptide \( \mathbf{A} \) from those that are noise or generated from peptide \( \mathbf{B} \). Given a spectrum \( \mathbf{M} \), the filtration stage then computes match scores between all candidate database peptide and their corresponding projected version of \( \mathbf{M} \); the top scoring matches are then selected for scoring of peptide-pair matches as described below. Because the projections extract peaks belonging to only one peptide from the query spectrum, the score between a projected spectrum and a peptide can be computed using any scoring function designed for single-peptide spectra; here we used the scoring function proposed by Kim et al. (25).

The efficiency of the projected-spectrum filter is determined by the highest (i.e. worst) rank of a correct peptide match of a mixture spectrum to the database \( \mathcal{D} \). Note that a correct match in \( \mathcal{D} \) is a match to one of the peptides generating \( \mathbf{M} \) - single-peptide spectra have one correct match, mixture spectra have two correct matches and thus two correct-match ranks. As shown in Fig. 1, the resulting ranks of correct matches indicate that the projected-spectrum filter is an efficient filter that, for over 96% of cases, retains both correct matches (i.e. maximum ranks) at ranks less than 500 from about 10,000 candidate peptides (yeast database, 3 Da precursor mass filtering). The left panel in Fig. 1 also shows that one of the correct matches, presumably the higher-abundance peptide in the mixture, almost always has rank less than 10 (i.e. minimum ranks). This means that for almost all cases one only needs to pair the top 10 candidates...
with the top 500 candidates to find the correct match. Using this strategy at most 10 \times 500 = 5000 candidate pairs need to be considered, thus conferring a \frac{10,000 \times 10,000}{2 \times 5000} \approx 10^4 speedup compared with considering all \approx 5 \times 10^7 possible candidate peptide pairs.

Scoring Function for Mixture Spectrum Matches—Efficient filtration dramatically reduces the numbers of candidate peptide pairs that one needs to consider but these candidate pairs must then be scored and ranked according to their relative likelihood of generating each putative mixture MS/MS spectrum. Although scoring a peptide against an MS/MS spectrum is a well studied problem in proteomics, few scoring functions have been designed to handle more than one peptide versus mixture MS/MS spectrum. Because each mixture spectrum has two correct matches we report both the minimum (i.e. best) and maximum (i.e. worst) rank of the two matches in the left and right panels, respectively. As shown, in more than 96% of cases, one of the correct matches has rank \leq 10 (left) whereas the other correct match ranks \geq 500 (right). Thus it is sufficient to pair the top 10 candidates with the top 500 candidates to find the correct pair. Using this strategy at most 10 \times 5000 peptide pairs need to be considered, resulting in a speedup of four orders of magnitude compared with considering all \approx 5 \times 10^7 possible peptide pairs.

Fig. 1. Filtration efficiency. Cumulative distributions of minimum rank (left) and maximum rank (right) for correct matches of simulated mixture spectra to the yeast database. Candidate peptides in the database are first sorted according to decreasing score against the corresponding projected mixture spectra (typical number of peptide candidates after precursor mass filtering is \approx 10,000). The ranks of the correct matches are then determined. Correct matches are peptides in the database that correspond to one of the peptides used to generate the simulated mixture spectrum. Because each mixture spectrum has two correct matches we report both the minimum (i.e. best) and maximum (i.e. worst) rank of the two matches in the left and right panels, respectively.

Because the additive contribution of each peak to the total peptide score is expressed by the log-ratio of probabilities of observing a peak rank for a particular ion type versus noise, such peak-rank models are usually very robust to the presence of low intensity noise peaks (see supplemental Fig. S1 for estimated rank distributions for noise and b/y ions). Conversely, high intensity noise peaks would cause the rank distributions for true ion peaks to shift toward worse ranks and, as expected, make them less distinguishable from high intensity noise; this would be reflected in lower peptide scores by resulting in lower log \left( \frac{\text{Prob}(s|p)}{\text{Prob}(s|0)} \right) scores for matched ion peaks. The values of \text{Prob}(s|p) can be learned from a training data set of annotated single-peptide spectra; similarly the noise model \text{Prob}(s|0) can be trained using the rank distribution of unassigned peaks in the same annotated single-peptide spectra. The learning is done separately for peptides of different precursor charge and length to account for their different fragmentation statistics (see (25) for full details of this model).

In order to score mixture spectrum matches, we extend this model for pairs of peptides (P_A P_B). Without loss of generality we refer to the highest-abundance peptide in the pair as P_A. To represent a pair of peptides in our statistical framework we extend the set of possible ion types to distinguish ions from P_A and those from P_B. If the original set of ion types I contained m elements I = \{i_1, i_2, \ldots, i_m\}, the extended ion types now have 2m elements: I_{\text{extended}} = I_A \cup I_B = \{i_A^{1}, i_A^{2}, \ldots, i_A^{m} \} \cup \{i_B^{1}, i_B^{2}, \ldots, i_B^{m}\}, where i_A^{j} represents the ion type \ i_j \ in \ P_A \ and \ i_B^{j} \ also \ represents \ the \ ion \ type \ associated \ with \ peptide \ P_B.
Peptide Identification by Database Search of Mixture Tandem Mass Spectra

TABLE I
Sensitivity of selecting the correct pair of peptides from the spectral library and protein sequence database. MixDB and M-SPLIT were evaluated on a set of 5000 simulated mixture spectra. Each row indicates the percentage of cases when the top ranking pair is correct; numbers in parenthesis indicate fraction of cases when one of the top ten peptide pairs is correct. Two types of database search were performed. In the first, each spectrum was searched against all Yeast tryptic peptides. In the second, each spectrum was searched only against peptides with spectra in the NIST Yeast spectral library. In general M-SPLIT spectral library search was found to have better performance than MixDB database search. However, as shown in the fourth column, the main advantage of spectral library search was the reduction in the number of possible peptide candidates. We also compared the MixDB’s performance with an iterative strategy where: (1) the dominant peptide is identified; (2) peaks explained by the top peptide match are removed from the spectrum, and (3) the residual spectrum is again searched against the database to identify the second peptide. When the mixture coefficient $\alpha$ is high the iterative strategy has similar performance to MixDB; however, as $\alpha$ becomes smaller, MixDB’s scoring of both peptides at the same time results in considerably better performance than the iterative strategy.

| Mixture coefficient ($\alpha$) | M-SPLIT | MixDB (whole Yeast database) | MixDB (library peptides only) | Iterative method |
|--------------------------------|---------|-----------------------------|------------------------------|-----------------|
| 1:1                            | 97      | 87 (97)                     | 95 (98)                      | 81              |
| 1:0.5                          | 92      | 79 (92)                     | 90 (98)                      | 74              |
| 1:0.3                          | 80      | 66 (86)                     | 79 (92)                      | 57              |
| 1:0.2                          | 63      | 50 (77)                     | 69 (87)                      | 30              |
| 1:0.1                          | 34      | 19 (43)                     | 34 (70)                      | 6               |

The performance of this scoring model is first evaluated using a set of simulated mixture spectra generated from a different data set than that used to generate the training data set. Using single-peptide spectra identified by both InsPecT and M-SPLIT in a previous study (18), we simulated mixture spectra with $\alpha = 1.0, 0.5, 0.3, 0.2, 0.1$ as described above. The percentage of cases where the top peptide pair returned by MixDB is correct is shown in Table I. As expected, as $\alpha$ decreases the performance worsens because it becomes harder to identify the lower abundance peptide. We also compared the performance of MixDB with M-SPLIT, our spectral library search tool previously shown to be efficient and robust in identifying mixture spectra. On average, M-SPLIT correctly identifies 15% more mixture spectra than MixDB. In general, spectral library search methods have two main advantages over database search methods: the relative intensities of different fragment ions are known in advance and the number of peptide candidates is smaller. In order to understand the relative importance of these two factors, we also evaluated MixDB when searching only against peptides with spectra in the NIST spectral library. As can be seen from Table I, with a reduced search space, MixDB has similar performance to M-SPLIT. We also compare MixDB with an iterative search strategy where one first identifies the highest-scoring peptide, removes all annotated peaks from the spectrum and then uses the “residual” spectrum to search against the database a second time to identify the second peptide. As we can see in Table I when $\alpha$ is high, the performance of the iterative method is comparable with that of the combined scoring function. However, as $\alpha$ becomes smaller, it is better to consider both peptides at the same time.

Classification of Database Search Matches—A database search of MS/MS spectra will always identify some top-scoring peptide or peptide pair for any given query spectrum, even if the true match is not in the database. To assess whether the top match is significant we use a two-stage classifier to distinguish true matches from false positive matches. Because our goal is to build a general search tool that can identify both single-peptide and mixture spectra we consider three possible outcomes when searching a given query spectrum $S$:

- No-match: $S$ does not match any peptide in the database.
- Single-peptide match: $S$ matches one peptide in the database.
- Mixture match: $S$ matches a pair of peptides in the database.

Classification of the top matches is done using two Support Vector Machines (SVMs) (26). The first SVM distinguishes No-match cases from Single-peptide/Mixture matches and the second SVM distinguishes Single-peptide matches from Mixture matches (see Fig. 3). To build the SVMs we consider the PPSM score described above and several other features that have been found useful in...
distinguishing true matches from false positives in single-peptide spectra, namely:
1) Likelihood score for one peptide match: likelihood score while considering only matched peaks from one peptide.
2) Likelihood score divided by peptide length: score from (1) divided by the number of amino acids in the top candidate peptide.
3) Explained intensity: total intensity of annotated peaks divided by total intensity of the spectrum.
4) Fraction of b and y ions present (two features): number of b and y ions present in the spectrum divided by the number of b/y ions possible from the peptide.
5) Longest consecutive series of b and y ions (two features).
6) Average mass error between theoretical and observed masses.

Features 2–6 are computed separately for each peptide in the top pair, resulting in a total of 16 feature inputs to the SVMs To train the SVM models, we constructed two negative control data sets. The No-match data set consists of 5000 mixture spectra where the peptides used to create the mixture spectra are deleted from the database. The Single-match data set consists of 5000 single-peptide spectra and 2500 mixture spectra where one peptide in the mixture is removed from the database. These two data sets were combined with another data set of 5000 mixture spectra (Mixture-match data set) and searched against the database. For all simulated mixture spectra, the mixture coefficient \( \alpha \) was selected uniformly from 0.1 to 1. The top matches from each data set were used as training data for the SVM models. The training is carried out in a two-step fashion. In the first step, no matches are considered by the second SVM as candidate mixture matches. The performance of the SVMs was assessed using cross-validation.

![Fig. 2. Classification of database matches.](image)

**RESULTS**

To illustrate the utility of our method in a typical scenario, we tested our database search method on an experimental *Yeast data set* (27), generously made publicly available in Tranche/ProteomeCommons (28) by researchers at Vanderbilt University. In brief, a tryptic digest of *Saccharomyces cerevisiae* was analyzed on an LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific) and MS/MS spectra were acquired using a data-dependent scanning mode in which each full MS scan (m/z 300–2000) was acquired on the Orbitrap at resolution 60,000, followed by eight MS/MS scans collected on the LTQ (see (27) for full details). The data were analyzed using InsPect (24), MiXDB, and ProbIDtree (17) with 3 Da parent mass tolerance and 0.5 Da fragment mass tolerance against the SGD yeast protein database (ver.5/8/2009); a 1% false discovery rate was enforced using a target and decoy strategy and the only modification allowed was car-
Peptide Identification by Database Search of Mixture Tandem Mass Spectra

**Fig. 3. Workflow of SVM classification and FDR estimation.** Every query spectrum searched against the database becomes assigned to some peptide pair \((P_a, P_b)\) that best matches the spectrum. Depending on whether the matched peptides come from the Target or Decoy databases, each match falls in one of four categories: Target/Target (TT), Target/Decoy (TD), Decoy/Target (DT), Decoy/Decoy (DD). All matches are ranked according to their SVM1 score to assess whether the most abundant peptide \((P_a)\) in each paired match \((P_a, P_b)\) is significant. As with the standard Target/Decoy strategy, matches with \(P_a\) from Target are considered positive matches and matches with \(P_b\) from Decoy are considered negative matches. Therefore, the FDR for single-peptide matches is computed as: 

\[
FDR_{match} = \frac{DD}{TD + DT + TT}
\]

and the SVM1 score is thresholded to yield 

\[
FDR_{match} < 0.01.
\]

Matches with SVM1 score above the threshold are then ranked by their second SVM scores to evaluate whether the second peptide \((P_b)\) in the top match \((P_a, P_b)\) is also significant. As before, matches \(P_b\) from Target database are considered positive matches and those with \(P_b\) from Decoy are negative matches. Thus the FDR for mixture matches is computed as:

\[
FDR_{mixture} = \frac{TD}{TT}
\]

and SVM2 scores are also thresholded such that 

\[
FDR_{mixture} < 0.01.
\]

boxamidomethylation on cysteine. Although data collected with this survey scan setting typically features very high mass accuracy, we allowed a large precursor mass tolerance in the searches to find possible MS/MS spectra from co-eluting peptides. In short, InsPecT is able to identify a total of 22,658 single-peptide spectra, MixDB is able to identify 23,930 single-peptide spectra plus 978 mixture spectra, and ProbiDtree is able to identify 19,840 single-peptide spectra plus 821 mixture spectra. Because the Yeast data set was acquired with high-accuracy survey scans, this information was further used to validate the annotations by requiring the presence of the theoretical monoisotopic \(m/z\) value of the identified peptides in the corresponding survey scans. Because for low-abundance peptides the monoisotopic \(m/z\) may not be visible in the corresponding survey scan (observed in only \(\sim 0.7\%\) of all cases and only for the lower abundance peptide in mixture spectra), in such cases we also check one preceding and one subsequent survey scan for the monoisotopic \(m/z\). An annotation is considered correct if the theoretical precursor \(m/z\) is within 5 ppm of the observed \(m/z\); the results are summarized in Table II. We first focus our attention on single-peptide cases. As seen in Table II, all three database search methods achieve similar precision of \(\sim 97\%\). A more detailed comparison in Fig. 4 shows that 78.4% and 68.2% of MixDB’s annotations overlap with those of InsPecT and ProbiDtree, respectively. For all spectra for which both MixDB and Inspect/ProbiDtree make an annotation, more than 96% of them have the same peptide ID, indicating that these independent methods are consistent. Among those spectra that are only identified by MixDB or InsPecT/ProbiDtree, we further divide them into two categories—those where two methods identify the same peptide as the top hit and those where the two methods have a different top hit. Those in the “same-top-hit” category are more likely to be correct, because different scoring functions rank them as the same top peptide candidate. If we consider these cases, it increases MixDB’s overlap with InsPecT and ProbiDtree to 90 and 85%, respectively, further indicating very good agreements between these different methods. Overall, for identification of single-peptide spectra, all three database search methods have comparable accuracy whereas MixDB identifies \(\sim 6\%\) more spectra than InsPecT and 21% more spectra than ProbiDtree. This shows that MixDB’s performance in identifying single-peptide spectra was not diminished by trying to identify more than one peptide per spectrum.

For mixture spectra, MixDB identifies 978 spectra whereas ProbiDtree identifies a total of 821 spectra (see Table IIA). MixDB identified two peptides in each mixture spectrum, whereas ProbiDtree found 32 mixture spectra with more than two identified peptides. This indicates that even though more than two identifiable co-eluting peptides may appear in one MS/MS spectrum this is relatively rare. Thus by limiting mixture spectra to two peptides per spectra MixDB does not lose much sensitivity in peptide identification. Furthermore, MixDB compensates by identifying about 20% more mixture spectra than ProbiDtree \((978 \text{ versus} \ 821\ \text{mixture spectra})\). In addition, by limiting its search space MixDB is more accurate. As shown in Table IIA, MixDB achieves a precision of almost 96% whereas ProbiDtree has only \(\sim 90\%\) precision for mixture spectra. Two main reasons contribute to MixDB’s higher precision. First, MixDB only searches up to two peptides per spectrum (ProbiDtree attempts to look for up to eight pep-
Peptide Identification by Database Search of Mixture Tandem Mass Spectra

Table II

Search results on the Yeast data set for MixDB, M-SPLIT, InsPecT, and ProbIDtree. (a) Numbers of identified spectra (single-peptide and mixture) and unique peptides are compared. (b) To allow for identification of co-eluting peptides, all searches were run using 3 Da precursor mass tolerance. The accurate precursor mass information was then used a posteriori to estimate the precision of peptide identification by comparing the theoretical precursor m/z of peptides returned by each method and the observed precursor m/z values in the corresponding MS1 scan (isotopic profile). An identification is considered correct if the difference between theoretical and experimental precursor m/z is less than 5 ppm. For mixture spectra the precision is slightly lower because the second peptide in the mixture is usually of low-abundance (average difference scores and FDRs for the first and second identified peptides per spectrum), thus theoretically it has a smaller search space than ProbIDtree. More importantly, MixDB applies different scores and FDRs for the first and second identified peptides in each MS/MS spectrum (FDRmatch and FDRmixture, respectively). This is crucial because high-abundance peptides are likely to have very different match statistics (e.g., % explained intensity, % b/y ions presented) than low-abundance peptides in mixture spectra. Because there are many more single-peptide spectra than mixture spectra in this data set, applying a single global FDR for the combined identification of both single-peptide and mixture spectra can lead to an underestimation of FDR for mixture spectra. Just as in the case of ProbIDtree, it has an estimated precision of 90.1% × 821 + 97.8% × 19840 / 19840 + 821 = 97.94% or an overall FDR of 2.06%. However, the precision for mixture spectra is only 90.1%, bringing its FDR on mixture spectra to 9.9% which is much higher than its FDR for single-peptide matches. In summary, we show that MixDB has both higher sensitivity and precision than ProbIDtree in identifying mixture spectra while having comparable performance to InsPecT in identifying single-peptide spectra.

Because spectral library methods are in general considered to be both more sensitive and more accurate in peptide identification than database search methods (29), we use identifications from spectral library searches to further validate the different database search methods. The Yeast data set was analyzed using an extension of the M-SPLIT algorithm (18) (See supplemental Methods) and SpectraST (30) with default parameters against the Yeast spectral library from NIST (ver.5/4/2009) with a precursor mass tolerance of 3 Da; a 1% false discovery rate was enforced using the decoy library strategy described in (31). Because we do not allow PTMs in the database searches we also removed all the entries in the spectral library that contain PTMs. To evaluate the performance of M-SPLIT, we first compare it to SpectraST, the most popular publicly available method for peptide spectral library search. In short we showed that both methods identified similar number of single-peptide spectra and are consistent with each other (see Supplemental Table S1). Only in ~6% of the spectra do the two methods identify different peptides as top hits (see supplemental Figure S2). Therefore, from here on we use results from M-SPLIT as a reference to evaluate results from different database search methods. As shown in Fig. 4, M-SPLIT misses about ~3600 single-peptide spectra identified by either MixDB or InsPecT. However, in 85–90% of these spectra, the identified peptides are not in the spectral library, indicating that M-SPLIT has very high sensitivity and therefore we can use spectra identified by M-SPLIT as a reference and compare the relative sensitivity of database search methods. Fig. 4 and supplemental Fig. S2 and S3 show that MixDB, InsPecT, and ProbIDtree identify 68%, 61.3%, and 51.8% of the spectra identified by M-SPLIT, respectively. Among these shared annotations more than 97% of them have the same peptide annotation as M-SPLIT, again indicating that these database search methods have high precision for single-peptide spectra. If we look further into those spectra that are only identified by M-SPLIT but not by database search methods, we find that for 60%, 30%, and 11% of the cases MixDB, InsPecT, and ProbIDtree, respectively, identify the same top hit as M-SPLIT. Altogether these indicate that MixDB, InsPecT, and ProbIDtree, have a sensitivity of 89%, 75%, and 58% for ranking the correct peptide as the top candidate. This implies that these database search methods are able to correctly identify these spectra, but the current scoring functions/SVM models do not have enough discriminative power to distinguish experiment-wide true
matches from false matches. Thus to keep the FDR low, database search methods have to discard some possibly valid but low-scoring matches.

Next, we turn our attention to mixture spectra. Similar to the single-peptide case, most mixture spectra identified by database search methods but missed by M-SPLIT corresponded to peptides that did not have spectra in the spectral library (see Fig. 4). Assuming the union of mixture spectra returned by M-SPLIT and MixDB as the total number of mixture spectra in this data set we get that M-SPLIT has a false negative rate of 
\[
\frac{2567 + 95 + 78}{78} = 0.0357
\]
at ranking the correct peptide pair as the top candidate and and 
\[
\frac{95}{2567 + 95} = 0.0274
\]
at the classification stage. Similar calculations for ProbIDtree results estimate the false negative rate of M-SPLIT as 0.056 and 0.0403 at ranking and classifying, respectively. This agrees with what we observed before (18) and again shows that M-SPLIT has high sensitivity and can also be used as a reference to compare database search methods for the identification of mixture spectra. Out of the 2567 mixture spectra identified by M-SPLIT, MixDB is able to identify 615 whereas ProbIDtree is able to identify 282. If we look at these cases closely, we find that about 2460 mixture spectra identified by M-SPLIT come from peptides with charge 2 and 3, the only possibilities considered by MixDB. This means MixDB has a sensitivity of 25% whereas probIDtree has a sensitivity of 12%. If the two peptides in mixture spectra are considered independent and observing that MixDB has a sensitivity of 60% for single-peptide spectra, we expect MixDB to identify 60% \* 60% = 36% of all mixture spectra identified by M-SPLIT. This calculation makes the assumption that the two peptides are present at similar abundance, which, as shown in the previous section, is the easiest scenario for identification of both peptides in mixture spectra. In practice, one peptide is present at lower abundance in most mixture spectra and in the yeast data set, the average mixture coefficient \(\alpha\) estimated by M-SPLIT is only 0.3. Thus having a sensitivity of 25% is a reasonable performance for MixDB. In addition, for a large fraction of mixture spectra (1051/1953) that MixDB did not classify as mixture matches, it identified the same top peptide pair per spectrum as M-SPLIT, but again the current SVM model does not have enough discriminative power to separate true matches from false positives. The extension of more sophisticated statistical models that have been proposed for single-peptide spectra (32) to mixture spectra is likely to increase the sensitivity of database search methods.

Fig. 4. Comparison of identifications from MixDB, M-SPLIT, InsPecT, and ProbIDtree. All pairwise comparisons between MixDB, M-SPLIT, InsPecT, and ProbIDtree compare identification results at 1% FDR as determined by the Target/Decoy strategy. In each pairwise comparison, spectra identified by both methods (in the intersection, shown in purple) were assigned to the same peptide in 96–97% of cases, indicating that the methods are consistent and the precision is in good agreement with our estimates. Spectra identified by one method but not the other are subdivided into two categories: cases where the two methods return the same peptide (peptide pair in the case of mixture matches) as the top hit but it was below the FDR threshold for one of the methods (shown in green) and cases where the two methods do not return the same peptide as the top hit (shown in black). In general MixDB has high overlap with other database search methods. For single-peptide spectra MixDB finds the same top peptide match as other methods in 85–90% of cases. When using spectra identified by spectral library search as a reference set, MixDB is able to identify 6–16% more single-peptide spectra and 38% more mixture spectra than current database search methods. Taken together, these show that MixDB has better and comparable sensitivity and accuracy in identifying single-peptide spectra as well as significantly higher sensitivity and accuracy in the identification of mixture spectra.
Peptide Identification by Database Search of Mixture Tandem Mass Spectra

In order to estimate how the presence of more than one peptide affects current computational techniques in peptide identification, we performed a more detailed comparison between MixDB and InsPecT on the 2567 mixture spectra identified by M-SPLIT. As described above, the problem of peptide identification consists of two subtasks: (1) ranking the correct peptide as the top scoring candidate per spectrum among all the other peptide candidates in the database and (2) experiment-wide discrimination of correct peptide matches from false, but top scoring matches. We evaluated how the presence of more than one peptide affected each of these tasks. First we compared whether InsPecT and MixDB are able to correctly rank one of the peptides in the mixture as the correct top match. As shown in Fig. 5, InsPecT is able to rank one of the correct peptides as the top match in 75% of cases, whereas MixDB was able to rank one of the peptides as the top match in 95% of cases and ranked both correct peptides as the top match in 65% of cases, showing that MixDB gains 20% higher sensitivity in spectrum identification by relaxing the assumption that each spectrum comes from only one peptide. After imposing score thresholds at 1% FDR, InsPecT was able to classify 53% of cases as single-peptide matches, whereas MixDB was able to classify 24% of cases as mixture matches and an additional 43% of cases as mixture matches. Because each mixture spectrum contains information for two peptides, InsPecT recovers 27% = 53% × 0.5 and MixDB recovers 46% = 24% + 43% × 0.5 of the peptide information contained in mixture spectra. 0.95.

DISCUSSION

As increasingly more complex samples are analyzed in high-throughput proteomic experiments (10) and new data acquisition protocols evolve (13, 14, 15, 33), the occurrence and detectability of co-eluting peptides per MS/MS spectrum is likely to increase. The almost ubiquitous assumption that mixture spectra contain more information than single-peptide spectra. In our analysis of a yeast data set, M-SPLIT, MixDB, and ProblIdtree identified 5997, 5476, and 4420 unique peptides from 28417, 23930, and 19840 single-peptide spectra, respectively. However, they were able to identify 2394, 1128, and 820 unique peptides from 2567, 978, 821 mixture spectra, thus revealing MixDB’s rate of 1128/978 = 115% peptide IDs per mixture spectrum versus 5476/23930 = 23% peptide IDs per single-peptide spectrum. These results show that if dynamic range challenges can be addressed (e.g. by selectively isolating precursors of comparable abundance (33)) then protocols designed to generate mixture spectra have the potential to substantially improve peptide identifications while considerably decreasing the number of MS/MS scans required to obtain these identifications. In addition, because of the lower-abundance nature of most of second IDs from mixture spectra, they are less likely to be sampled again by the instrument and may not even be detectable without a co-eluting peptide. Altogether, methods that consider more than one peptide per MS/MS spectrum can potentially double the sensitivity in recovering peptide information from mixture MS/MS spectra.

However, this higher information content comes at a cost in that the combinatorial explosion of searching for best peptide pairs dramatically increases the search space for mixture spectra. Thus, effective filtration strategies and special con-
Peptide Identification by Database Search of Mixture Tandem Mass Spectra

consideration for controlling the FDR become crucial in achieving high precision in the identification of mixture spectra. Similar to the development of any mass spectrometry algorithm, a large data set of reliably identified spectra is crucial and is often hard to come by. In this study we addressed this issue with comprehensive simulations and by taking advantage of the public availability of the yeast spectral library (23) and experimental data (27). Spectral library search methods are in general considered to be more sensitive and accurate for peptide identification (29) and in this case the yeast spectral library is also comprehensive (i.e. only a small fraction of peptides identified by database search method are not in the library). Thus we can use identifications from spectral library searches as the reference “truth” and comprehensively benchmark various aspects of different database search methods. We showed that MixDB has good sensitivity and high precision in identifying both single-peptide spectra and mixture spectra.

The development of computational methods and novel experimental strategies often rely on each other. For example, because mainstream computational approaches assume each MS/MS spectra comes from one peptide, development of chromatography and mass spectrometry protocols has focused on making this assumption valid for most cases. However, with a new generation of algorithms that are able to identify mixture spectra, experimental protocols designed to generate mixture spectra may lead to many interesting applications. There are already several alternative data acquisition approaches that rely on mixture spectra to overcome the limitations of current instrument scan rates (13, 14, 15). Mixture spectra also arise from peptides that are covalently linked in the sample, examples include disulfide bridges (34), SUMOylated peptides (35), and peptides from cross-linking experiments (36, 37). The identification of these cross-linked peptides can provide valuable information on protein structures and interactions. Thus solving the problem of peptide identification from mixture spectra represents an important step toward addressing related and emerging problems in proteomics.

Acknowledgments—We thank NIST, ProteomeCommons, and Vanderbilt University for the public availability of the mass spectrometry data used in this research.

This work was supported in part by the National Institutes of Health grant GM078596 (JW and PEB) and 1-P41-RR024851 (NB) from the National Center for Research Resources.

REFERENCES

1. Bonneil, E., Tessier, S., Carrier, A., and Thibault, P. (2005) Multiplex multidimensional nanoLC-MS system for targeted proteomic analyses. Electrophoresis 26, 4575–4589
2. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., and Whitehouse, C. M. (1989) Electrospray ionization for mass spectrometry of large biomolecules. Science 246, 64–71
3. McLuckey, S. A., Goeringer, D. E., and Glish, G. L. (1992) Collisional activation with random noise in ion trap mass spectrometry. Anal. Chem. 64, 1455–1460
4. Washburn, M. P., Wolters, D., and Yates, J. R. 3rd (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat. Biotechnol. 19, 242–247
5. Brunner, E., Ahrens, C. H., Mohanty, S., Baetschmann, H., Loevenich, S., Potthast, F., Deutsch, E. W., Panse, C., de Lichtenberg, U., Rinner, O., et al. (2007) A high-quality catalog of the Drosophila melanogaster proteome. Nat. Biotechnol. 25, 576–583
6. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. Nature 422, 198–207
7. Chalkley, R. J., Baker, P. R., Huang, L., Hansen, K. C., Allen, N. P., Rexach, M., and Burlingame, A. L. (2005) Comprehensive analysis of a multidimensional liquid chromatography mass spectrometry data set acquired on a quadrupole selecting, quadrupole collision cell, time-of-flight mass spectrometer. Mol. Cell. Proteomics 4, 1194–1204
8. Werner, B. R., and Lynn, B. C. (2004) Factors that affect ion trap data-dependent MS/MS in proteomics. J. Am. Soc. Mass Spectrom. 15, 150–157
9. Alves, G., Ogurtsov, A. Y., Kwok, S., Wu, W. W., Wang, G., Shen, R. F., and Yu, Y. K. (2008) Detection of co-eluted peptides using database search methods. Biol. Direct 3, 27
10. Michalski, A., Cox, J., and Mann, M. (2011) More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent i? ions. J. Proteome Res. 10, 1785–1793
11. Luethy, R., Kessner, D. E., Katz, J. E., MacLean, B., Grothe, R., Kani, K., Fa?a, V., Pitteri, S., Hanash, S., Agus, D. B., and Mallick, P. (2008) Precursor-ion mass re-estimation improves peptide identification on hybrid instruments. J. Proteome Res. 7, 4031–4039
12. Houel, S., Abernathy, R., Renganathan, K., Meyer-Arendt, K., Ahn, N. G., and Old, W. M. (2010) Quantifying the impact of chimera ms/ms spectra on peptide identification in large-scale proteomics studies. J. Proteome Res. 9, 4152–4160
13. Venable, J. D., Dong, M. Q., Wohlschlegel, J., Dillin, A., and Yates, J. R., 3rd. (2004) Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. Nat. Methods 1, 39–45
14. Masselon, C., Pasa-Tolic, L., Lee, S. W., Li, L., Anderson, G. A., Harkewicz, R., and Smith, R. D. (2003) Identification of tryptic peptides from large databases using multiplexed tandem mass spectrometry: simulations and experimental results. Proteomics 3, 2624–2633
15. Chakraborty, A. B., Berger, S. J., and Geblert, J. C. (2007) Use of an integrated MS-multiplexed MS/MS data acquisition strategy for high-coverage peptide mapping studies. Rapid Commun. Mass Spectrom. 21, 272–280
16. Geiger, T., Cox, J., and Mann, M. (2010) Peptide identification on an orbitrap bench-top mass spectrometer using all-ion fragmentation. Mol. Cell. Proteomics 9, 2252–2261
17. Zhang, N., Li, X. J., Ye, M., Pan, S., Schwickowski, B., and Aebersold, R. (2005) ProtiDtree: an automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. Proteomics 5, 4096–4106
18. Wang, J., Pérez-Santiago, J., Katz, J. E., Mallick, P., and Bandeira, N. (2010) Peptide identification from mixture tandem mass spectra. Mol. Cell. Proteomics 9, 1476–1485
19. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal. Chem. 74, 5383–5392
20. Elias, J. E., and Gygi, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat. Methods 4, 207–214
21. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S., and MacCoss, M. J. (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat. Methods 4, 923–925
22. Bern, M., Cai, Y., and Goldberg, D. (2007) Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tan-

10.1074/mcp.M111.010017–10
31. Lam, H., Deutsch, E. W., and Aebersold, R. (2010) Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics. J. Proteome Res. 9, 605–610
32. Kim, S., Gupta, N., and Pevzner, P. A. (2008) Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. J. Proteome Res. 7, 3354–3363
33. Michalski, A., Damoc, E., Hauschild, J. P., Lange, O., Wieghaus, A., Makarov, A., Nagaraj, N., Cox, J., Mann, M., and Horning, S. (2011) Mass spectrometry-based proteomics using x exact, a high-performance benchtop quadrupole orbitrap mass spectrometer. Mol. Cell. Proteomics
34. Choi, S., Jeong, J., Na, S., Lee, H. S., Kim, H. Y., Lee, K. J., and Paek, E. (2010) New algorithm for the identification of intact disulfide linkages based on fragmentation characteristics in tandem mass spectra. J. Proteome Res. 9, 626–635
35. Pedrioli, P. G. A., Raught, B., Zhang, X. D., Rogers, R., Aitchison, J., Maturis, M., and Aebersold, R. (2006) Automated identification of SU-MOylation sites using mass spectrometry and SUMmOn pattern recognition software. Nat. Methods 3, 533–539
36. Schilling, B., Row, R. H., Gibson, B. W., Guo, X., and Young, M. M. (2003) MS2Assign, automated assignment and nomenclature of tandem mass spectra of chemically crosslinked peptides. J. Am. Soc. Mass Spectrom. 14, 834–850
37. Rinner, O., Seebacher, J., Walzthoeni, T., Mueller, L. N., Beck, M., Schmidt, A., Mueller, M., and Aebersold, R. (2008) Identification of cross-linked peptides from large sequence databases. Nat. Methods 5, 315–318

In order to cite this article properly, please include all of the following information: Wang, J., Bourne, P. E., and Bandeira, N. (2011) Peptide Identification by Database Search of Mixture Tandem Mass Spectra. Mol. Cell. Proteomics 10(12):M111.010017. DOI: 10.1074/mcp.M111.010017.