MixGF: spectral probabilities for mixture spectra from more than one peptide

Jian Wang[1], Philip E. Bourne[2], Nuno Bandeira[2,3,4]

[1]Bioinformatics Program, University of California, San Diego, La Jolla, USA

[2]Skaggs School of Pharmacy and Pharmaceutical Sciences, UCSD, San Diego, La Jolla, USA

[3]Center for Computational Mass Spectrometry, University of California, San Diego, La Jolla USA

[4]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, USA

*Correspondance to: bandeira@ucsd.edu; phone: 1-858-534-8666; fax: 1-858-534-7029

Running Title: Spectral probability for mixture tandem mass spectra Correspondence to:

Nuno Bandeira

Center for Computational Mass Spectrometry

Department of Computer Science and Engineering

University of California, San Diego

9500 Gilman Drive, Mail Code 0404

La Jolla, CA 92093-0404, USA

Email: bandeira@ucsd.edu

Phone: 1-858-534-8666

Abbreviations

- MS/MS: Tandem mass spectrometry
- 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
- PRM: Prefix Residue Mass
- mPSM: multi-Peptide Spectrum Match
- DDA: Data Dependent Acquisition
- DIA: Data Independent Acquisition
- TDA: Target Decoy Approach

Supplementary Materials

Estimation of False Discovery Rates

False Discovery Rate (FDR) can be estimated by extending the Target-Decoy Approach (TDA) for database search [1]. Each top scoring peptide-peptide-spectrum match (PPSM) can be of one of the following types: *TT* – both peptide matches are from the target database, *TD* or *DT* – one peptide is from the target while the other peptide is from the decoy database and *DD* – both peptides are from the decoy database. A peptide from the target database can be either a correct (C) or incorrect match (I) and a peptide from the decoy database is by definition an incorrect match. Therefore matches of each type can be further divided into subtypes: for example, TT can be divided into TT^{CC}, TT^{CI}, TT^{IC}, and TT^{II} where the superscript indicates whether the peptide match is correct or not. We can then write the number of PPSMs belonging to each type as a sum of PPSMs belonging to its subtypes:

$$\begin{aligned} TT &= TT^{CC} + TT^{CI} + TT^{IC} + TT^{II} && (1)\\ TD &= TD^{CI} + TD^{II} && (2)\\ DT &= DT^{IC} + DT^{II} && (3)\\ DD &= DD^{II} && (4) \end{aligned}$$

TDA assumes that an incorrect peptide match has equal chance of coming from the target or the decoy database. Therefore matches of type II has equal chance of being TT, TD, DT and DD, making the number of II matches in equation 1-4 approximately the same: $DD^{II} = DT^{II} = TD^{II} = TT^{II}$. By a similar argument, the number of matches of type CI in equation 2 and the number of matches of type IC in equation 3 should be comparable to those in equation 1. Hence the extension of TDA to mixture spectra is made using the following equivalences derived from the standard TDA assumptions that matches to decoy are appropriate models for the false matches to target:

$$DD = DD^{II} = DT^{II} = TD^{II} = TT^{II} \tag{5}$$

$$TT^{CI} = TD^{CI} \tag{6}$$

$$TT^{IC} = DT^{IC} \tag{7}$$

To test whether these hold true, we constructed a set of simulated mixture spectra (see next section) and extracted the top-scoring matches of type II, CI and CI returned by MixDB and compute the relative frequency of each peptide match being from the target or decoy database. As shown in Figure S1a, matches of type II had $\sim 25\%$ chance of being TT, TD, DT, and DD. Figure S1b shows that within range of random variation, matches of type CI has equal probability of being TT and TD while Figure S1c shows that matches of types IC, has equal chance of being TT and DT. Taken together, these results show that the TDA assumption can be generalized to mixture spectra.

By substitution and rearranging terms, we can thus redefine the CI and IC terms in equation 1 to:

$$TT^{CI} = TD - DD \tag{8}$$

$$TT^{IC} = DT - DD \tag{9}$$

As described in the Methods section, two different FDRs need to be computed for MixGF: one is used to determine the probability threshold for the joint probability and the other for determining the threshold for conditional probability. Joint probability aims to accept PPSMs that are of type TT^{CC}, TT^{CI}, TT^{IC} and reject matches of the type TT^{II}. Thus we are interested in controlling the following FDR: $FDR_{Joint} = \frac{TT^{II}}{TT^{CC}+TT^{CI}+TT^{IC}+TT^{II}}$. Conditional probability aims to accept matches of type TT^{CC} and reject matches of the other types. As such, the FDR of the conditional probability can then be defined as: $FDR_{Cond} = \frac{1/2(TT^{IC}+TT^{CI})+TT^{II}}{TT}$ where the $1/2$ in the equation accounts for the fact that matches of IC and CI type contribute one correct match and one incorrect match. Substituting the terms defined above, we get:

$$FDR_{Joint} = \frac{DD}{TT} \tag{10}$$

$$FDR_{Conditional} = \frac{1/2((TD-DD)+(DT-DD))+DD}{TT} \tag{11}$$

$$= \frac{1/2(TD+DT)}{TT} \tag{12}$$

Testing the TDA assumption for mixture spectra

In order to test whether the TDA assumptions can be extended to mixture spectra, we generated a set of simulated mixture spectra of type CI, IC and II using the NIST spectral libraries [2] for Yeast and E. coli where the E. Coli spectral library was pruned to remove any entries assigned to peptide sequences matching Yeast protein sequences. Mixture spectra of type II were simulated by linearly combining two spectra from the E.Coli library while mixture spectra of type CI and IC were simulated by linearly combining one spectrum from the Yeast library and one spectrum from the E. Coli library. All the simulated mixture spectra were searched against a Yeast protein sequence database using MixDB [3]. The top-scoring match of type II. CI and IC were extracted and used to compute the frequency that each match was from TT, TD, DT and DD respectively. Each experiment was performed on a set of three thousand simulated mixture spectra (one thousand for each type: II, CI, CI) and the experiment was repeated twenty times to estimate the random fluctuations in the data (shown as error bars in Figure S1).

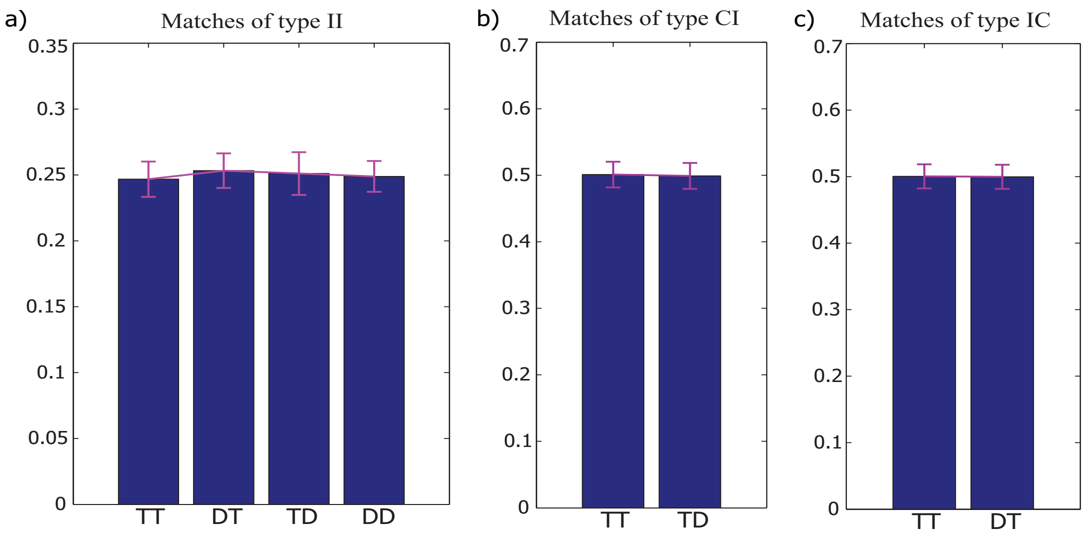

Figure S1: Target/Decoy Approach (TDA) for mixture spectra: TDA assumes that for an incorrect peptide-spectrum-match (PSM), the peptide has equal chance of coming from the target (T) or the decoy (D) database. For a peptide-peptide-spectrum match (PPSM) there are three kinds of incorrect matches: II– where both peptides are incorrect and CI/IC – where one peptide is a correct match and the other peptide is an incorrect match. Extending the TDA assumption to mixture spectra implies that a match of type II will have equal chance of being TT, TD, DT or DD. Similarly a match of type CI will have equal chance of being TT and TD while a match of type IC will have equal chance of being DT and TT. In order to test these assumptions, we constructed a set of simulated mixture spectra and extracted the top-scoring matches of type II, CI, and IC returned by MixDB and computed the relative frequency of each peptide being from the target (T) or the decoy (D) database. As shown in a)-c), within the range of random variations, each incorrect peptide match (I) has approximately 50% chance of being from the target or the decoy database, confirming that the TDA assumption can be generalized to mixture spectra as proposed.

Testing MixGF on higher-complexity mixture spectra

In order to test how the complexity of mixture spectra affects MixGF's performance we constructed sets of simulated mixture spectra from two to five peptides and searched them with MixGF. Simulated spectra were created in a similar fashion as for the two-peptide case (see Method). Briefly, to create a mixture spectra with K peptides, K single-peptide spectra: $S_1, S_2, ...S_k$ with precursor mass difference less than 3 Da were selected randomly from the spectral library. The sum of peak intensity of each selected spectra were first normalized to one. A simulated mixture spectrum, M, was then created by linearly combining the single-peptide spectra: $M = S_1 + \alpha_2 S_2 ... \alpha_k S_k$. The mixture coefficient α_i which reflects relative abundance of the peptide in the mixture spectrum was randomly chosen between 0.3 to 1.0. MixGF's sensitivity at identifying the first two peptides at 1% FDR were shown in Table ST1. It is observed that MixGF is robust at identifying higher-complexity mixture spectra as the sensitivity of identifying the top two peptides showed no significant changes as more peptides were added in the mixture spectra.

Number of peptides in mixture spectra	2	3	4	5
MixGF sensitivity at identifying top two peptides	80.0	83.7	80.3	77.5

Table ST1: MixGF sensitivity for mixture spectra with higher complexity: Simulated mixture spectra that are from two to five peptides were generated and analyzed by MixGF. The sensitivity at identifying the top two peptides by MixGF were shown.

MixGF for high mass accuracy data

The dynamic programming approaches proposed in the main text to compute spectral probabilities for mixture spectra assumes that amino acids have integer masses, thus resulting in significant rounding errors for high mass accuracy MS/MS data. The following text described the extensions introduced in the implementation of MixGF to take advantage of high mass accuracy.

Spectrum processing

All MS/MS peaks were first deconvoluted and moved to their monoisotopic mass position as previously described in [4]. The spectra were then filtered to retain only the 12 most intense peaks in each window of 50 Da.

Mass error model for PRM spectrum

As mentioned in the paper, a PRM spectrum is an transformation of an observed spectrum to a scored version of the spectrum that can be used to compute a score for a PSM or PPSM. However, since mass values for each PRM spectrum are discretized to 1 Da bins, theoretical fragment ions from peptides do not match peaks in the spectrum with the small mass errors characteristics of high accuracy instruments. To utilize the high mass accuracy information in MS/MS data, we used a similar approach to that introduced in MSGF+ [5] where the main idea is that if MS/MS peak masses are accurate then mass *differences* between pairs of MS/MS peaks should also be proportionally accurate when the paired peaks come from consecutive ions of the same ion type. This can be illustrated using an example: when matching the peptide $LARGER$ to a spectrum S, peaks m_4 and m_5 are matched to the $b4$ and $b5$ fragment ions respectively (i.e,. right before and right after E). If $LARGER$ is the correct peptide match, the mass difference, $\delta = mass(m_5) - mass(m_4)$ should be close to the mass of the amino acid E with any mass error expected to be within twice the mass error tolerance of the instrument. Conversely if $LARGER$ is an incorrect peptide match to S, then m_4 and m_5 are random matches and thus it is unlikely their mass difference will be close to $mass(E)$.

To extend the model using this idea, we define a fragment ion type as a pair $I = (o, d)$, where o is the mass offset of the ion type and d denotes whether the ion is a prefix ion or suffix ion (note that the charge of an ion type is not considered here because peaks were first converted to charge 1 by the preprocessing step). A prefix ion is a fragment ion that contain the N-termus of the peptide while a suffix fragment ion is an ion that contains the C-terminus of the peptide. For simplicity of notation, we also define a peak by its mass

m since peak intensity is not considered in the following discussion. Given an ion type $I_j = (o_j, d_j)$, for a particular peak m we can define its PRM position to be:

$$PRM(m, I_j) = \begin{cases} Round(m - o) & \text{if } d = Prefix \\ Round(Parentmass - (m - o)) & \text{if } d = Suffix \end{cases} \tag{13}$$

where $Round$ returns the nearest integer for a real number. Therefore, given an ion type I, each peak in the spectrum is assigned to a unique PRM position. Let x and y be a pair of PRM positions such that they differ by approximately the mass of an amino acid (without loss of generality we assume $x > y$):

$$(x, y)|x - y = a \in Round(Amino\ Acid\ Mass) \tag{14}$$

Given an ion type I, two peaks m_x^I and m_y^I form a pair if m_x^I is assigned to a PRM x and m_y^I is assigned to a PRM y such that x and y satisfy the condition above. We define a differential mass error for a peak pair as: $\delta(m_x^I, m_y^I, a) = m_x^I - m_y^I - a$. Let D be a random variable that represents the differential mass errors of all peak pairs in a set of PSMs. The probability distribution of D for peak pairs in correct PSMs, $Prob(D|true\_match)$, can be learned from a set of annotated spectra and similarly a background distribution $Prob(D|decoy\_match)$ for peak pairs from incorrect matches can also be learned from a set of decoy PSMs obtained by searching a set of spectra against a decoy database. A likelihood score can then be defined as:

$$ErrorScore(x, y, I) = log(\frac{Prob(D = \delta(m_x^I, m_y^I, a)|true\_match)}{Prob(D = \delta(m_x^I, m_y^I, a)|decoy\_match})) \tag{15}$$

which reflects the mass error for a peak pair. Note that if x or y have no assigned peak, δ is assigned a special value of ∞. The probability of the converse of this event: $Prob(D \neq \infty)$ reflects the chance of matching two consecutive ions of the same ion type in the spectrum and it is expected to be higher for correct matches than incorrect matches. Thus $D = \infty$ will result in a negative likelihood score whose actual value is learned from the data. Given a set of ion types $\{I_1, I_2 ... I_t\}$, a score can be calculated for every pair of PRM positions satisfying Equation 14 by summing up the error scores for each ion type:

$$ErrorScore(x, y) = \sum_{i=1..t} ErrorScore(x, y, I_i) \tag{16}$$

Thus for an MS/MS spectrum with parent mass M we can construct a PRM spectrum S of size M and an $M \times M$ matrix, E, where $E_{i,j}$ contains the $ErrorScore$ described above (note only entries $E_{i,j}|i - j \in AminoAcidMass$ in this matrix are valid). To score a peptide P with PRMs $P = p_1, p_2, ...p_n$ against the spectrum M, we add the score from both S and E: $Score(P, S) = \sum_{i=1..n} S_{p_i} + \sum_{i=1..n-1} E_{p_i,p_{i+1}}$. Using this scoring function the spectral probability can be computed using dynamic programming method similar to the one described in the main text.

Refining PRM spectrum for conditional probability

It is often found that in mixture spectra, MS/MS peaks used as ions for PRMs matched to the first peptide in a PPSM often interfere as high-intensity 'noise' from the perspective of the second peptide, thus effectively decreasing the statistical significance of the second peptide. To avoid this effect, we aim to remove these peaks from the spectrum before calculating the conditional probability. Given a peak m and a fragment ion annotation I, a PRM position can be computed as described above and the corresponding PRM score is used as a confidence score that the ion annotation I is correct for m. For example, if a peak m is assigned to a $b - H_2O$ ion in the first peptide, we use the score at M^H_{m+18} to assess how likely this annotation is to be correct. Similarly one can iterate through all possible ion types and use the PRM spectrum to determine what is the most likely ion annotation for every peak in the spectrum matched to the first peptide in the PPSM. The main idea is that m will most likely be from the first peptide if no better annotation from the second peptide can be found considering all possible ion types. More concretely, given a PRM spectrum M, the score for a peak m given an ion type I is: $Score(m|I_j, M) = M_{PRM(m,I_j)}$. For a mixture spectrum let m_i be a peak that is matched to the first peptide and I^i be the ion annotation for that peak. Also let $\mathcal{I}$ be the set of all possible fragment ion types that are being considered: $\mathcal{I} = \{I_1, I_2...I_n\}$. A peak m_i will be assigned to the first peptide if:

$$Score(m_i|I^i, M^H) \geq Score(m_i|I_j, M^L) \text{ for all} j = 1..n \quad (17)$$

All peaks from the first peptide matching equation 17 were removed from the spectrum. Then peak intensity ranks and the PRM spectrum M^L for the low-abundance peptide were recomputed and the conditional probability was determined as described in the main text.

References

[1] J E Elias and S P Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*, 4:207–214, 2007.

[2] Eds. S.E. Stein and P.A. Rudnick. NIST Peptide Tandem Mass Spectra LIbraries. Yeast Peptide Mass Spectral Reference Data, ion trap, 2009, National Institute of Standards and Technology, Gaithersburg, MD, 20899.

[3] J. Wang, P.E. Bourne, and N. Bandeira. Peptide identification by database search of mixture tandem mass spectra. *Molecular & Cellular Proteomics*, 10(12), 2011.

[4] Adrian Guthals, Karl R Clauser, and Nuno Bandeira. Shotgun protein sequencing with meta-contig assembly. *Molecular & Cellular Proteomics*, 11(10):1084–1096, 2012.

[5] Sangtae Kim and Pavel A Pevzner. Universal database search tool for mass spectrometry. *submitted for publication*.