Supplements of the Generating Function of CID, ETD and CID/ETD Pairs of Tandem Mass Spectra: Applications to Database Search

1 MS-GF training

MS-GF takes a set of peptide-spectrum matches (PSMs) as a training set and outputs a file containing scoring parameters. All spectra in the training set are assumed to be generated using the same fragmentation method and the same enzyme. Below we describe the following 5 steps for generating MS-GF scoring parameters: (1) partitioning the training set, (2) selecting precursor offsets for removal, (3) selecting ion types, (4) computing peak rank distributions and (5) computing peak error distributions. We remark that steps (2) and (3) were missing in the previous MS-GF version [1] thus forcing users to specify the ion types manually. Note that by adding the steps (2) and (3), MS-GF can now automatically learn scoring parameters from any type of spectra with any precursor charges.

1.1 Partitioning the training set

Fragmentation propensities of mass spectra strongly depend on the precursor charge [2] and the peptide length [3].[1] Therefore, we use different sets of scoring parameters depending on the precursor charge, and the peptide length. To generate the scoring parameters, we partition the training set by the precursor charge of the spectrum and the estimated peptide length inferred from the precursor mass of the spectrum. Then, for each partition, we learn the parameters using only spectra belonging to this partition. In addition, since different types of peaks have different propensities

[1] The differences in the fragmentation propensity between peptides of similar lengths (like 7 and 8 amino acids) are typically small as compared with differences between peptides with very different lengths (like 7 and 20 amino acids).

with respect to the relative positions in the spectrum (e.g. peaks corresponding to doubly charged ions only appear in the lower part of the spectrum), we learn the parameters separately for the lower and the upper halves of the mass range.

1.2 Selecting precursor offsets for removal

ETD spectra often possess precursor peaks, charge-reduced precursor peaks, their neutral losses and side-chain losses [4]. While those peaks usually have high intensities, they do not contribute useful information for peptide identifications. We therefore remove these peaks to avoid a risk of erroneously interpreting them as other ion types. To figure out which the peaks have to be removed, we use the offset frequency function (OFF) [5]. OFF is a histogram of the peaks observed at a relative offset from a specific m/z in the spectrum. Here we use the OFFs from the precursor mass and charge-reduced precursor m/z's.

First we filter all spectra in the training set to remove noisy peaks as follows: given a peak at mass m, we retain the peak if it is among the top k (k = 6 by default) peaks within a window of size 100 Da around m. Then we compute the OFFs from the precursor m/z and all possible charge-reduced precursor m/z's. If a certain offset is observed in more than a predefined portion of the spectra (15% by default), we mark the offset for removal. Later, all peaks observed at marked offsets are filtered out (see Figure 1).

1.3 Selecting ion types

For each partition of the training set, we select ion types to be used for scoring using the OFF of the prefix and suffix residue masses as described in Dancik et al., 1999 [5]. We represent an ion type by a triplet of (charge, prefix or suffix, offset) and consider all possible prefix and suffix ions with charges 1 to the precursor charge and integer offsets from -38 to +38. If we observe an ion type at more than a predefined portion of all cleavage sites in the filtered spectra (15% by default), we select the ion type.

1.4 Computing peak rank distributions

For each selected ion type at a certain partition, we compute the probability of a peak of rank i being the ion type (ion rank probability) from rank 1 to $MaxRank$ (150 by default). We also compute the probability of a peak of rank i being an ion type that is not selected (noise rank

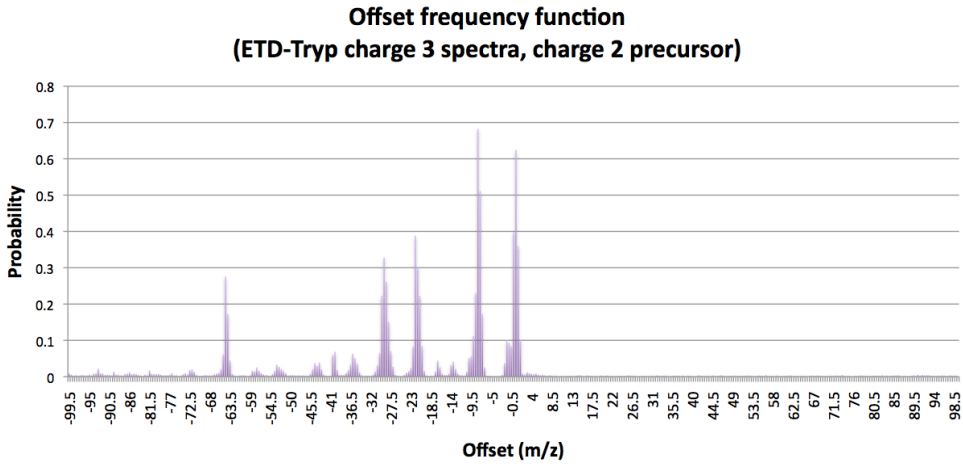


Figure 1: Example of the offset frequency function (OFF) from the (charge-reduced) precursor m/z. Shown is the OFF from the charge 2 (charge-reduced) precursor m/z of charge 3 spectra in ETD-Tryp data set. The horizontal axis represents the distance (in m/z) from the charge-reduced precursor m/z. The vertical axis represents the probability that a peak of the corresponding offset exists. For example, in about 40% of the charge spectra in ETD-Tryp data set, there exists a peak with m/z corresponding to the charge 2 precursor m/z minus 22. All the offsets over a predefined probability (0.15 by default) are marked for removal.

probability). As described in [3], the log of the ion rank probability over the noise rank probability at certain rank serves as the *rank score* of a peak.

1.5 Computing peak error distributions

Instead of setting up a fixed mass error threshold (e.g. 0.5 Da for ion traps) and assigning the same score to all peaks within this error threshold, we vary scores depending on the mass error. To do this, for each selected ion type at each partition, we compute the mass error histogram of all peaks of ranks within $MaxRank$ assigned to that ion type (ion error probability). We also compute a similar histogram using ion types that are not selected (noise error probability). The log ratio of the ion error probability over the noise error probability serves as the *error score* of a peak. See Supplement 2 and Figure 2 for how the rank and error scores are used to compute the Prefix-Residue-Mass (PRM) spectrum.

2 Review of the MS-GF scoring function

Here, we briefly review the MS-GF scoring function described in [3] and the generating function of tandem mass spectra described in [1]. MS-GF computes two different but related scores: the

MS-GF score and the p-value. The MS-GF score evaluates the quality of a PSM, and is used as a statistic to compute the p-value. The p-value (also termed "spectral probability" in [3]) of a PSM is the sum of the probabilities of all peptides with scores equal to or better than the MS-GF score of this PSM. Below we formally define these two scores.

Given a peptide $P = a_1 \dots a_n$ of mass m, we define its theoretical spectrum (denoted by $Spectrum(P)$) as a 0-1 vector $p_1 \dots p_m$ with $(n-1)$ 1s, such that $p_i = 1$ if i is the mass of the peptide $a_1 \dots a_i$ $(1 \leq i < n)$. Given a spectrum S of parent mass m, the Prefix-Residue Mass (PRM) spectrum of S (denoted by $PRMSpectrum(S)$) is defined as an integer vector of dimension m. The ith coordinate of the PRM spectrum represents the log likelihood ratio that the peptide from which the spectrum was derived contains a prefix of mass i (see Figure 2 for an example) [3, 5]. The MS-GF score (denoted as $MSGFScore(P, S)$) between a peptide P and a spectrum S is defined as the dot product of $PRMSpectrum(S)$ and $Spectrum(P)$ if the parent mass of S and P is the same, and $-\infty$ otherwise. Amino acid probabilities are predefined depending on the frequencies of amino acids in a protein database and the probability of a peptide is defined as the product of probabilities of its amino acids. Then, the p-value of a PSM between spectrum S and peptide P (denoted by $Pvalue(P, S)$) is defined as follows:

$$Pvalue(P,S) = \sum_{\text{all peptides } P' \text{ with } MSGFScore(P',S) \geq MSGFScore(P,S)} Probability(P').$$

P-values can be efficiently computed by the generating function approach [1].

3 Comparison of MS-GFDB with Mascot using spectrum-level FDR

in the main text, we showed the number of peptide identifications against peptide-level false discovery rates (FDRs). Here, we show the similar figures of the number of identified *spectra* against the spectrum-level FDRs (Figure 3).

4 MS-GFDB outperforms OMSSA, SEQUEST and iProphet on ETD-LysC data set.

in the main text, we have demonstrated that MS-GFDB outperforms Mascot for various types of spectra. Here we show the performance of MS-GFDB is also superior to other popular database

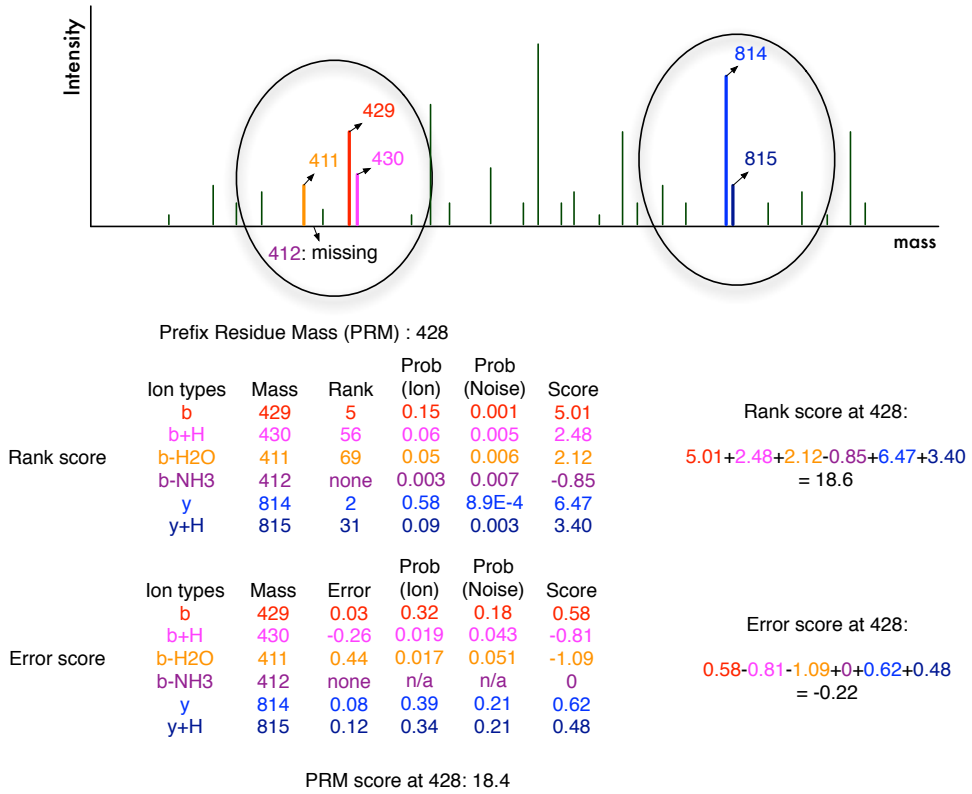


Figure 2: Example of PRM spectrum computation. For simplicity, only b, b+H, b-H_2O, b-NH_3, y and y+H ions are considered for scoring. The ion rank/error probabilities and noise rank/error probabilities are pre-computed from training data as described in Supplement 1. The rank/error score is defined as the log of the ion rank/error probability over the noise rank/error probability. The PRM score at a certain mass is defined as the sum of the rank score and the error score. For example, at prefix residue mass 428, there exist b, b+H, b-H_2O, y and y+H b, b+H ions in the spectrum so the rank/error score of each ion is given depending on the rank/error of the corresponding peak. b-NH_3 ion is missing, thus the negative rank score and zero error score is added. The PRM score at 428 is set as the sum of the rank score and the error score (18.4).

search tools such as SEQUEST [6] and OMSSA [7], and even to post-processing tools such as PeptideProphet [8, 9] and iProphet.[2]

Publicly available and previously described data set generated by the Coon lab [10] was used in the analysis. It is acquired from a yeast whole-cell lysate sample digested with Lys-C, prefractionated with SCX, and analyzed via LC-MS/MS on a custom LTQ Orbitrap outfitted with ETD capability. This data set is extensively analyzed by the Aebersold lab in the Institute of Systems Biology using ETD-adapted Trans Proteomic Pipeline (version 4.3) [11]. We downloaded the data set

[2]Shteynberg, D., et al., Postprocessing and validation of tandem mass spectrometry datasets improved by iProphet. in preparation.

including spectrum files, protein database files and search results of OMSSA, SEQUEST, PeptideProphet and iProphet from Tranche (http://proteomecommons.org/tranche/). To train MS-GFDB scoring model for ETD spectra of Lys-C peptides, we used 4,993 spectra of distinct peptides where the iProphet probabilities obtained by combining SEQUEST and OMSSA results exceeding 0.9. MS-GFDB was carried out against the same database with the same parameters as described in Deutsch et al.'s paper [11] .

We compared the performance of MS-GFDB with all other tools by counting the number of identified spectra for each FDR (spectrum-level FDR) using the combined TDA.[3] MS-GFDB outperformed OMSSA and SEQUEST, and even PeptideProphet and iProphet that re-score OMSSA and SEQUEST peptide-spectrum matches using multiple features (many of them are unavailable to MS-GFDB) to get more identifications (Figure 4). For example at 1% FDR, MS-GF identified 18,605 spectra, corresponding to 40%, 23%, 180%, 32% and 7% improvement over OMSSA, OMSSA+iProphet, SEQUEST, SEQUEST+iProphet and SEQUEST+OMSSA+iProphet.

5 MS-GFDB outperforms Percolator.

Here we show that MS-GFDB outperforms Percolator [12, 13, 14]. The four data sets (CID-Tryp, ETD-Tryp, CID-LysN and ETD-LysN) and the same human database described in the main text were used in the comparison. Since Mascot (version 2.3.0) automatically runs Percolator, we obtained percolator results reported by Mascot. Posterior error probabilities were used to compute FDRs.

Percolator significantly improves on Mascot (especially for ETD spectra from trypsin digests). However, it identified less peptides than MS-GFDB at the same FDR (Figure 5). For example, at 1% peptide-level FDR, while MS-GFDB identified 13,584/11,978/6,831/7,544 peptides, Percolator identified 13,003/11,421/4,602/4,017 peptides from the CID-Tryp/ETD-Tryp/CID-LysN/ETD-LysN data set.

[3]Instead of running programs ourselves, we used the the results of OMSSA, SEQUEST, PeptideProphet and iProphet reported in [11] and the combined TDA was the only available approach for FDR computation with the results.

6 Specific choice of the training data set does not significantly affect the MS-GFDB results.

To generate the results presented in the main text, we used the same data set for both training and evaluation of the performance, thus one may raise a valid question that the improved performance of MS-GFDB may be due to the over-fitting of parameters to the specific data set. Here we show this is not the case. For the ISB Lys-C data set (used in the Supplement 3), we executed MS-GFDB search twice, using two different scoring parameter files: one with parameters derived using the ISB Lys-C data set itself and the other with parameters derived using the ETD-Tryp data set (used in the main text). Note that two data sets are generated in different laboratories (one from the Coon lab and the other from the Heck lab). We measured the number of identified spectra for different FDRs with MS-GFDB using the two scoring parameter files and the difference was small (Figure 6). For example, at 1% FDR the difference in the number of identified spectra was only 3%. This indicates that the specific choice of the training data set does not significantly affect the MS-GFDB results.

7 Using OMSSA to analyze CID/ETD pairs

OMSSA allows users to specify ion types to be used for scoring and can potentially use the following strategy to analyze CID/ETD pairs: for each CID/ETD pair, construct a new spectrum by merging all peaks in both spectra and run OMSSA allowing for b, y, c and z ions. We applied this strategy for the analysis of the CID-Tryp, ETD-Tryp, CID-LysN and ETD-LysN data sets (Figure 7). The results clearly show that this strategy leads to inferior results; across the entire FDR range, OMSSA identified fewer peptides by using "merged" spectra than using only CID spectra. For example, at 1% peptide level FDR, OMSSA CID/ETD (using merged spectra) identified 9,330 and 3,267 peptides whereas OMSSA CID (using only CID spectra) identified 11,251 and 4,093 peptides from trypsin and Lys-N digests, respectively.

8 Comparison of MS-GFDB and Mascot in the identification of phosphorylated peptides.

MS-GFDB can be used to analyze modified (e.g. phosphorylated) peptides. To demonstrate this, we acquired a data set generated with the same protocol as described in the main text but from a sample enriched for phosphorylated peptides using strong cation exchange [15]. The data set is composed of 68,670 CID/ETD spectral pairs (33,463 from trypsin digests and 35,207 from Lys-N digests).

Both Mascot and MS-GFDB were run with the following parameters: 50 ppm precursor mass tolerance, 0.5 Da fragment ion tolerance, up to 2 missed cleavages allowed, carbamidomethyl Cys as fixed modification and phosphorylation of Ser, Thr and Tyr as variable modifications. We used the same scoring parameters of MS-GFDB as we described in the main text.

Figure 8 shows the number of identified peptides against the corresponding FDRs. Peptides with the same sequence and the same type/number of modifications but different modification sites were counted as one. MS-GFDB identified more peptides than Mascot across the entire FDR range. For example, at 1% peptide level FDR, MS-GFDB identified 1130, 916, 787 and 817 peptides while Mascot identified 880, 560, 560 and 439 peptides from CID spectra of trypsin digests, ETD spectra of trypsin digests, CID spectra of Lys-N digests and ETD spectra of Lys-N digests, respectively. Similar trends were observed when we counted the number of phosphorylated peptides only. At 1% FDR, MS-GFDB identified 794/706/653/670 phosphorylated peptides and Mascot identified 602/434/525/408 phosphorylated peptides from CID spectra of trypsin digests/ETD spectra of trypsin digests/CID spectra of Lys-N digests/ETD spectra of Lys-N digests.

References

[1] Kim, S., Gupta, N., Pevzner, P.: Spectral probabilities and generating functions of tandem mass spectra: A strike against decoy databases. J. Proteome Res. **7** (2008) 3354–63

[2] Huang, Y., Triscari, J.M., Tseng, G.C., Pasa-Tolic, L., Lipton, M.S., Smith, R.D., Wysocki, V.H.: Statistical characterization of the charge state and residue dependence of low-energy cid peptide dissociation patterns. Anal. Chem. **77** (2005) 5800–13

[3] Kim, S., Gupta, N., Bandeira, N., Pevzner, P.: Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra. Mol. Cell. Proteomics **8** (2009) 53–69

[4] Savitski, M.M., Nielsen, M.L., Zubarev, R.A.: Side-chain losses in electron capture dissociation to improve peptide identification. Anal. Chem. **79** (2007) 2296–302

[5] Dancík, V., Addona, T.A., Clauser, K.R., Vath, J.E., Pevzner, P.A.: De novo peptide sequencing via tandem mass spectrometry. J. Comput. Biol. **6** (1999) 327–42

[6] Eng, J., McCormack, A., Yates, J.: An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. **5** (1994) 976–89

[7] Geer, L.Y., Markey, S.P., Kowalak, J.A., Wagner, L., Xu, M., Maynard, D.M., Yang, X., Shi, W., Bryant, S.H.: Open mass spectrometry search algorithm. J. Proteome Res. **3** (2004) 958–64

[8] Choi, H., Nesvizhskii, A.I.: Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. J. Proteome Res. **7** (2008) 254–65

[9] Keller, A., Nesvizhskii, A., Kolker, E., Aebersold, R.: Empirical statistical model to estimate the accuracy of peptide identifications made by ms/ms and database search. Anal. Chem. **74** (2002) 5383–92

[10] Swaney, D.L., McAlister, G.C., Coon, J.J.: Decision tree-driven tandem mass spectrometry for shotgun proteomics. Nat. Methods **5** (2008) 959–64

[11] Deutsch, E.W., Shteynberg, D., Lam, H., Sun, Z., Eng, J.K., Carapito, C., von Haller, P.D., Tasman, N., Mendoza, L., Farrah, T., Aebersold, R.: Trans proteomic pipeline supports and improves analysis of electron transfer dissociation datasets. Proteomics **10** (2010) 1190–95

[12] Käll, L., Canterbury, J.D., Weston, J., Noble, W.S., Maccoss, M.J.: Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat. Methods **4** (2007) 923–5

[13] Käll, L., Storey, J.D., Maccoss, M.J., Noble, W.S.: Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. J. Proteome Res. **7** (2008) 29–34

[14] Brosch, M., Yu, L., Hubbard, T., Choudhary, J.: Accurate and sensitive peptide identification with mascot percolator. J Proteome Res **8** (2009) 3176–81

[15] Gauci, S., Helbig, A.O., Slijper, M., Krijgsveld, J., Heck, A.J.R., Mohammed, S.: Lys-n and trypsin cover complementary parts of the phosphoproteome in a refined scx-based approach. Anal. Chem. **81** (2009) 4493–501

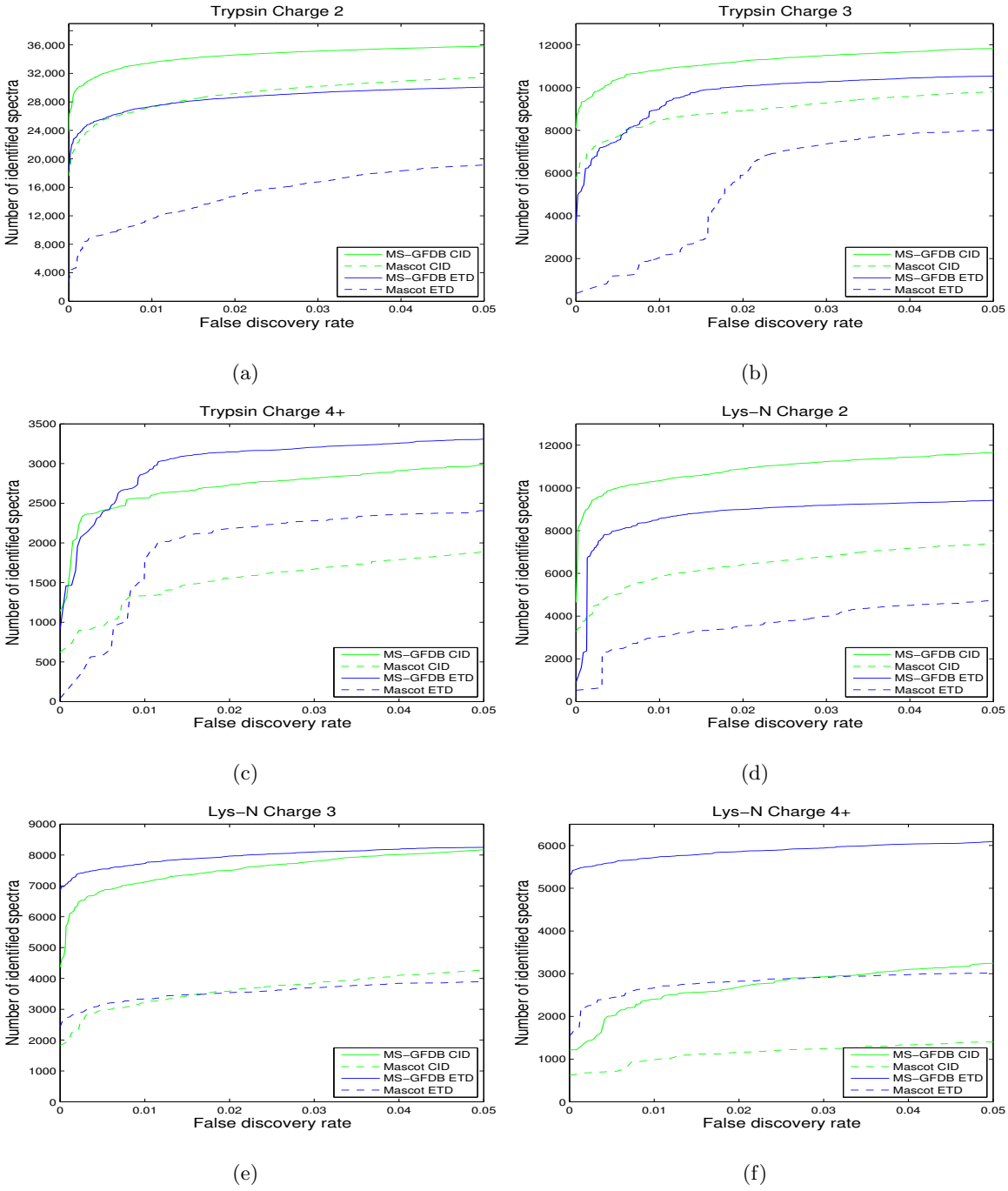


Figure 3: Number of identified spectra with Mascot and MS-GFDB from (a) charge 2 spectra in CID-Tryp and ETD-Tryp, (b) charge 2 spectra in CID-LysN and ETD-LysN, (c) charge 3 spectra in CID-Tryp and ETD-Tryp, (d) charge 3 spectra in CID-LysN and ETD-LysN, (e) spectra with charges 4 and larger in CID-Tryp and ETD-Tryp, and (f) spectra with charges 4 and larger in CID-LysN and ETD-LysN. The number of identified spectra in the IPI-Human database is plotted against the corresponding false discovery rate. Solid curves represent MS-GFDB and dashed curves represent Mascot. Green curves represent CID and blue curves represent ETD. Mascot ion scores and MS-GFDB p-values were used for computing FDR. For all the cases considered, MS-GFDB outperformed Mascot.

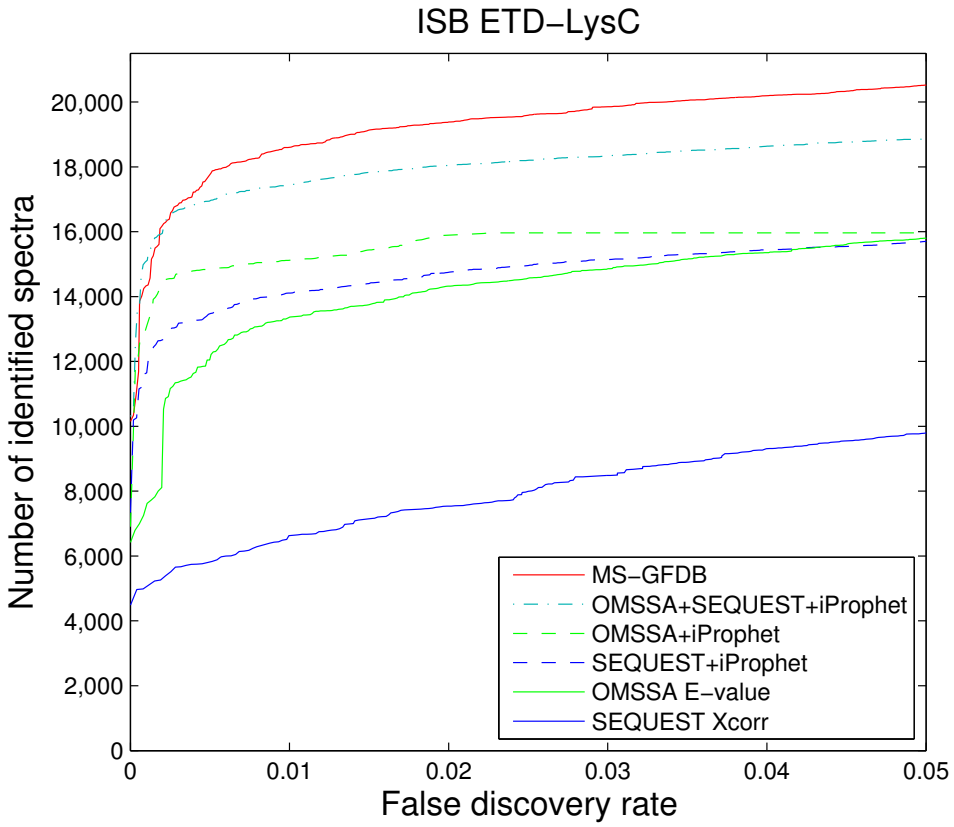
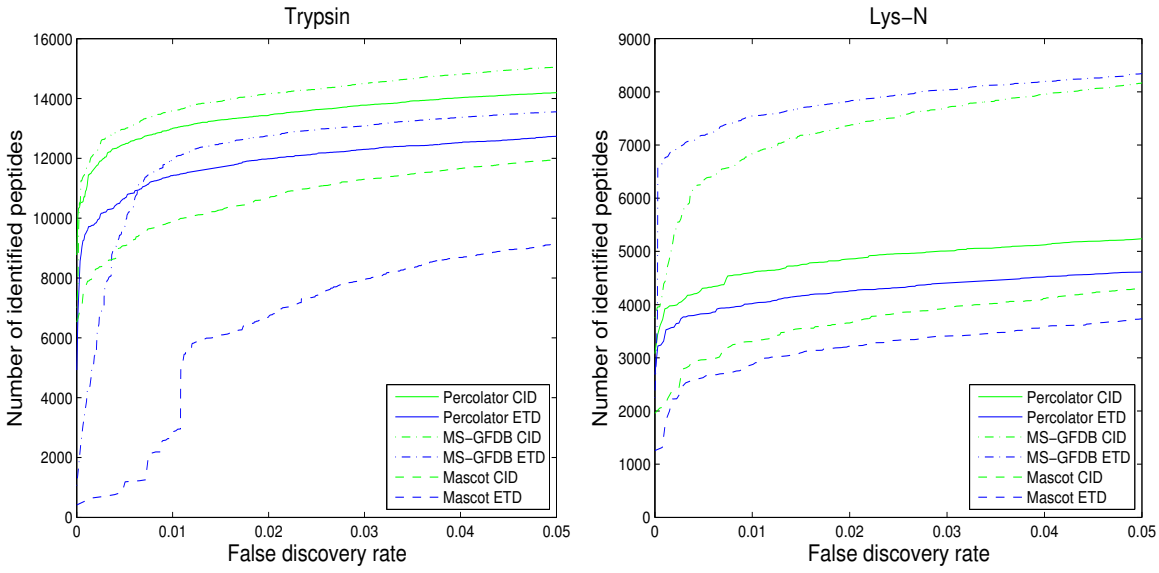


Figure 4: Number of identified spectra (out of 61,020 spectra) with SEQUEST, OMSSA, OMSSA+iProphet, SEQUEST+iProphet, OMSSA+SEQUEST+iProphet and MS-GFDB. False discovery rates were calculated using the combined TDA. OMSSA+PeptideProphet and SEQUEST+PeptideProphet show similar results as OMSSA+iProphet and SEQUEST+iProphet, respectively and thus are not shown. MS-GFDB outperformed all other tools.



(a) (b)

Figure 5: Number of identified peptides with Percolator at varying FDRs from the (a) trypsin digests (b) Lys-N digests. MS-GFDB and Mascot results were also shown for reference. Posterior error probabilities were used to compute FDRs. Percolator significantly improved on Mascot (especially for ETD spectra from trypsin digests), but identified less peptides than MS-GFDB at the same FDR.

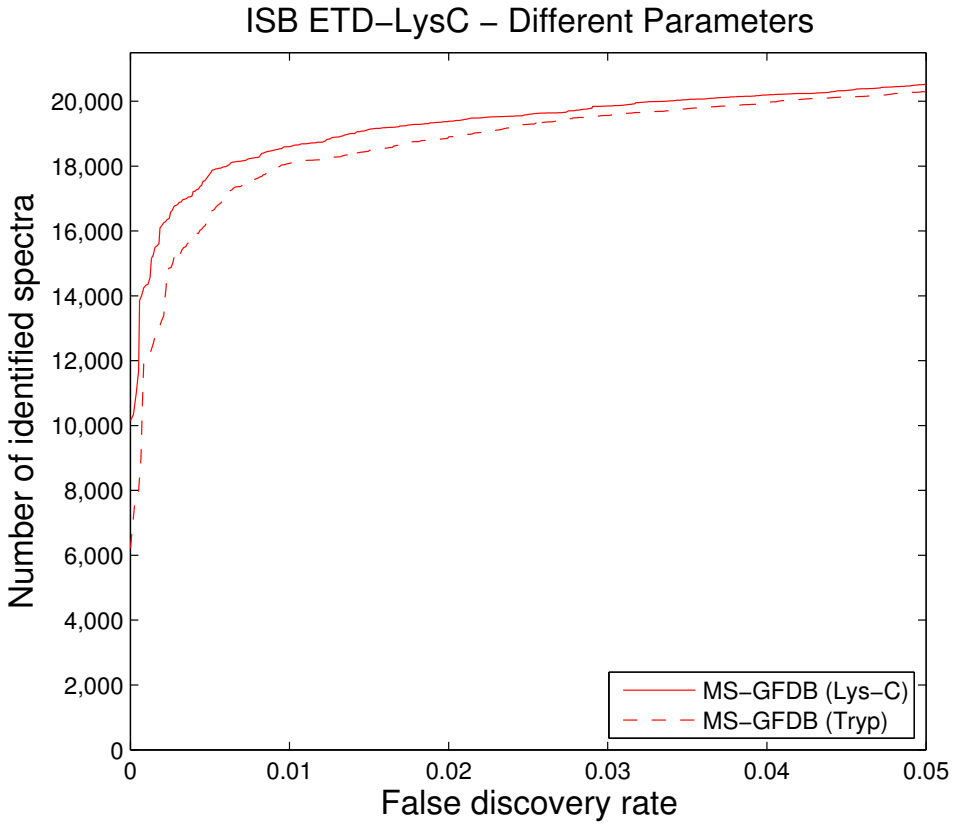
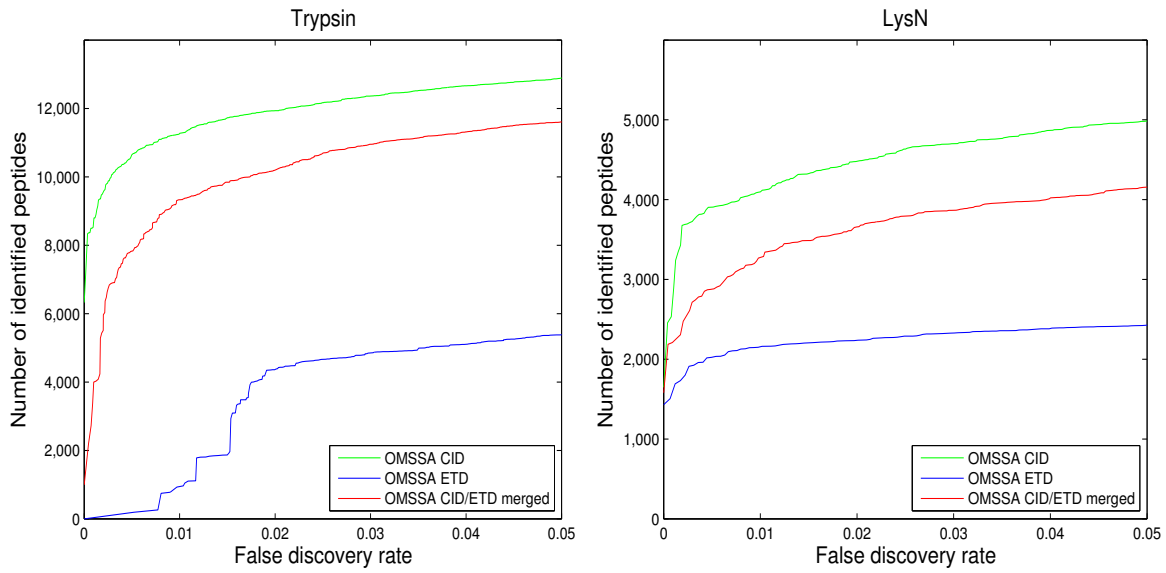
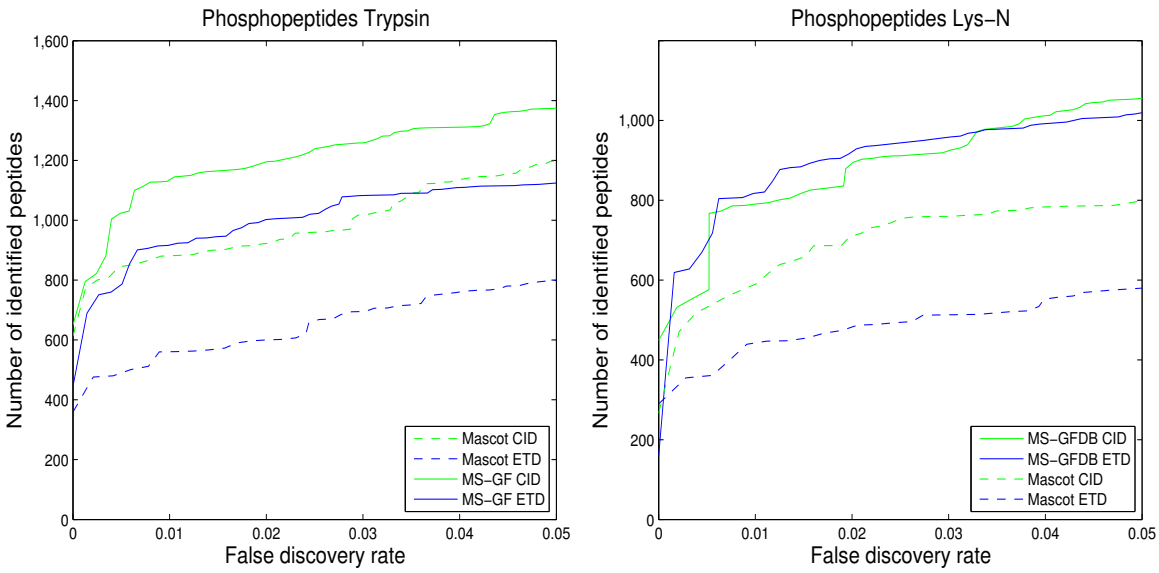


Figure 6: Number of identified spectra in the ISB Lys-C data set with MS-GFDB, using the scoring parameters derived from the ISB Lys-C data set from the Coon lab (solid line) and ETD-Tryp data set from the Heck lab (dashed line). The specific choice of the training data set does not significantly affect the MS-GFDB results



(a) (b)

Figure 7: Number of identified peptides with OMSSA at varying FDRs from the (a) CID-Tryp and ETD-Tryp data sets and (b) CID-LysN and ETD-LysN data sets. OMSSA CID (or ETD) represents the identifications using only CID (or ETD) spectra in the OMSSA search by enabling b and y ions (or c and z ions). OMSSA CID/ETD represents the identifications using the merged CID/ETD spectra in the OMSSA search by enabling b, y, c and z ions. OMSSA CID/ETD identified less number of peptides than OMSSA CID across the entire FDR range.



(a) (b)

Figure 8: Number of identified peptides with Mascot and MS-GFDB from the phosphorylation enriched data set with (a) trypsin digests and (b) Lys-N digests. The number of identified peptides in the target database is plotted against the corresponding false discovery rate. Solid curves represent MS-GFDB and dashed curves represent Mascot. Green curves represent CID and blue curves represent ETD. Mascot ion scores and MS-GFDB p-values were used to compute the FDR. For all the four data sets, MS-GFDB identified more peptides than Mascot at the same FDR.