Published in final edited form as:

J Proteome Res. 2008 November ; 7(11): 4878–4889. doi:10.1021/pr800484x.

Adaptive Discriminant Function Analysis and Re-ranking of MS/MS Database Search Results for Improved Peptide Identification in Shotgun Proteomics

Ying Ding^{1,2}, Hyungwon Choi^{1,2}, and Alexey I. Nesvizhskii^{1,3}

¹Department of Pathology, University of Michigan, Ann Arbor, MI 48109

²Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109

³Center for Computational Biology and Medicine, University of Michigan, Ann Arbor, MI 48109

Abstract

Robust statistical validation of peptide identifications obtained by tandem mass spectrometry and sequence database searching is an important task in shotgun proteomics. PeptideProphet is a commonly used computational tool that computes confidence measures for peptide identifications. In this paper, we investigate several limitations of the PeptideProphet modeling approach, including the use of fixed coefficients in computing the discriminant search score and selection of the top scoring peptide assignment per spectrum only. To address these limitations, we describe an adaptive method in which a new discriminant function is learned from the data in an iterative fashion. We extend the modeling framework to go beyond the top scoring peptide assignment per spectrum. We also investigate the effect of clustering the spectra according to their spectrum quality score followed by cluster-specific mixture modeling. The analysis is carried out using data acquired from a mixture of purified proteins on four different types of mass spectrometers, as well as using a complex human serum dataset. A special emphasis is placed on the analysis of data generated on high mass accuracy instruments.

Keywords

Tandem Mass Spectrometry; Database searching; Peptide Identification; Statistical Modeling; Adaptive Discriminant Analysis; Mass Accuracy; Decoy Sequences

INTRODUCTION

Tandem mass spectrometry (MS/MS) followed by sequence database searching is the method of choice for high-throughput identification of proteins in complex biological samples^{1, 2}. Proteomics is increasingly dependent on automated large scale MS/MS database queries using SEQUEST³, Mascot⁴, X! Tandem⁵ and similar tools^{6–8}. The statistical methods for validation of peptide assignments to MS/MS spectra play a central role in the entire mass spectrometry (MS) data analysis pipeline^{6, 9, 10}.

PeptideProphet¹¹ is a commonly used statistical validation tool which converts MS/MS database search score(s) into a probability that peptide identification is correct. Originally developed to analyze SEQUEST search results, it has since been extended to other search

tools, including X! Tandem and Mascot. More recent improvements combined the probabilistic modeling of PeptideProphet with the target-decoy strategy (e.g. addition of decoy sequences to the searched protein sequence database) into a semi-supervised mixture modeling framework¹². This led to improved robustness and higher accuracy of computed probabilities of correct identifications in the case of low quality datasets where unsupervised mixture modeling may not be able to accurately capture the underlying distributions of scores. The parametric assumptions of the original implementation can also be relaxed via a recently described semi-parametric mixture modeling approach¹³.

There are several limitations of the PeptideProphet method that have yet to be carefully investigated. First, in the case of SEQUEST and X! Tandem search tools, prior to unsupervised or semi-supervised mixture modeling, the model uses a set of fixed weighting factors to combine several database search scores into a single discriminant search score. These coefficients were derived using spectra produced on an LCQ ion trap mass spectrometer from a control sample of known protein components¹¹ and may not be optimal under all circumstances^{14, 15}. Another limitation is the use of the top-ranked (based on the primary search score) peptide assignment for each MS/MS spectrum only. In some cases, the top-ranked peptide is incorrect, but the correct one is not far below in the ranking. Hence, utilizing the primary score-based top-ranked peptide only may cause a non-ignorable loss of correct peptide assignments that otherwise can be extracted from the data. Although it is possible to use multiple peptide assignments per spectrum in the mixture modeling implementation of Scaffold¹⁶, and in several other recently described tools^{14, 15}, the benefits of doing this have not been investigated in detail. Finally, questions remain regarding the validity of modeling all spectra in the dataset using one common mixture distribution even though the spectra may differ significantly in terms of their spectral quality¹⁷.

In this paper, we investigate these limitations and describe a dynamic modeling approach to address them. In particular, instead of using the fixed weighting coefficients in computing the discriminant search score, we employ an adaptive method in which a new discriminant function is learned from the data. We also extend the modeling framework to go beyond the top scoring peptide assignment per spectrum. To achieve this, the top 10 peptides (based on the primary score) are taken for each spectrum and then re-ranked within each spectrum using the computed posterior probabilities to derive the new top-ranked peptide. Finally, we investigate the effect of clustering the spectra according to their spectrum quality score followed by cluster-specific mixture modeling. The analysis is carried out using several datasets of varying complexity, from control protein mixtures to a human serum sample, and generated using different mass spectrometers. The database search program SEQUEST is used as a representative search tool. A special emphasis is placed on the analysis of data generated on high mass accuracy instruments, and in particular the role of mass tolerance and other database search constraints in the validation process.

METHODS

Experimental Data

Protein Mix Data—This dataset was taken from Ref.¹⁸. The MS/MS spectra were collected using a mixture of purified trypsin digested proteins ('protein mix 3') on four different mass spectrometers: Thermo LTQ-FT, Thermo LTQ, Agilent XCT Ultra, and Waters/Micromass QTOF Ultima. The spectra were searched using SEQUEST against a database containing the sequences of proteins known to be present in the sample appended with a much larger database of reversed human protein sequences. The searches were performed in the enzyme semi-constrained mode (tryptic cleavage at least at one of the termini), and allowing at most one missed internal cleavage site, and using large mass tolerance window (3 Da, average mass; denoted 'LW'). A fixed modification of 57.02 was

specified for cysteine residues. The spectra collected using the LTQ-FT mass spectrometer were additionally searched in four other modes: narrow mass tolerance (0.025 Da, monoisotopic mass; denoted 'NW'), enzyme unconstrained ('Unconst'), semi-tryptic ('Semi'), or tryptic ('Tryptic') searches, and large mass tolerance (3 Da, monoisotopic), enzyme unconstrained search. All matches to peptide sequences from the sample proteins were considered to be correct, whereas all matches to peptides derived from reversed human protein sequences were considered incorrect¹¹.

Human Serum—This dataset was generated by the Western Consortium of the National Cancer Institute's Mouse Models of Human Cancer¹⁹. The sample was processed through the Multiple Affinity Removal System (MARS) to deplete albumin, transferrin, IgG, IgA, anti-trypsin, and haptoglobin from human serum. Only one LC-MS/MS replicate (file MARS_humansera_01) was used in this work. The spectra were searched using SEQUEST against the IPI human protein database version 3.32 appended with an equal number of reversed protein sequences. The search was done using 3 Da mass tolerance, average mass, semi-tryptic search, and allowing at most one missed internal cleavage site. A fixed modification of 57.02 was specified for cysteine residues.

Database search scores and discriminant function

Many search tools compute multiple scores useful for discrimination between correct and incorrect identifications. These scores together with a variety of peptide properties are used to assess the validity of peptide assignments. The primary database search score (e.g. Xcorr score in SEQUEST) measures the similarity between the acquired MS/MS spectrum and a theoretical spectrum constructed for each peptide in the searched protein sequence database. In addition to the primary score, related measures can be defined, e.g., the delta score measuring the relative distance between the scores of the top and the second best database matches for a given spectrum^{11, 20–25}. To simplify the subsequent mixture modeling, in PeptideProphet multiple database search scores are combined into a single discriminant score^{11, 12}.

Specifically, with SEQUEST, three search scores are used^{11, 12}: 1) Xcorr', the log-transformed and length-normalized cross-correlation (Xcorr) score, 2) ΔCn , the normalized difference between the Xcorr scores for the best and the second best scoring peptides, and 3) SpRank, which measures how the top scoring peptide ranked with respect to other candidate peptides during the preliminary scoring step, log transformed. The discriminant score, denoted S , is calculated through the following equation

$$S = c_1 Xcorr' + c_2 \Delta Cn + c_3 \ln(SpRank) + c_4 \quad (1)$$

with the coefficients in Eq. 1 determined using the linear discriminant analysis (LDA). LDA is a method that projects the multiple search scores onto a real line which best separates correct from incorrect identifications. The SEQUEST discriminant score coefficients were trained using a dataset generated on a mixture of 18 trypsin digested proteins with spectra collected using an LCQ ion trap instrument, enzyme unconstrained search, large mass tolerance (3 Da). The spectra of different charge states are modeled separately.

Posterior probabilities and FDR via mixture modeling

The database search scores represent only one set of discriminant features. Other useful features can be defined⁶ that are related to the digestion step (e.g., the number of tryptic termini and missed cleavages, NTT and NMC), peptide separation step (e.g. the difference between the observed and predicted retention time), or the first stage of the MS measurement (e.g. dM, the difference between the measured and calculated peptide mass),

see Figure 1 for illustration of the whole process. Additional discriminant parameters (e.g. *pI* value) can be defined when appropriate depending on the experimental protocol.

The joint distribution of the discriminant search score *S* and auxiliary peptide information (in this work NTT, NMC, dM), collectively denoted *D*, is modeled as a multivariate mixture distribution with two components representing correct and incorrect identifications respectively. It is assumed that conditional on the identification status, the marginal distributions of the individual variables are independent. The modeling is performed using the expectation maximization (EM) algorithm, and requires the prior specification of the shapes of the distribution for the discriminant search score *S*. Based on empirical observations using training data, the distribution of *S* in the case of SEQUEST are modeled with a Normal distribution and a Gamma distribution for correct and incorrect identifications, respectively, for detailed description see Ref.¹¹. The probability that the identification is correct given the discriminant search score and peptide properties can be calculated as

$$p = p(\text{Correct}|D) = \frac{\pi_1 f_1(D)}{\pi_1 f_1(D) + \pi_0 f_0(D)} \quad (2)$$

with $\pi_0 + \pi_1 = 1$.

The estimation of the parameters of the distributions among correct and incorrect identifications, $f_1(D)$ and $f_0(D)$, and the mixing proportions, π_0 and π_1 can be either performed by the regular EM algorithm (*unsupervised* EM), or using the recently described *semi-supervised* mixture modeling framework¹². The semi-supervised modeling can be carried out when for some of the peptide identifications the class label (correct or incorrect) is known, e.g., by appending decoy protein sequences to the searched sequence database²⁶. In those cases, decoy matches are incorporated in the estimation via modified overall log likelihood, with decoys contributing to the estimation of incorrect distribution only. Furthermore, although not used in this work, the parametric assumption can be relaxed via semi-parametric mixture modeling¹³.

The probabilities computed using Eq. 2 can be interpreted as the complement of the local false discovery rate^{12, 27–29}. They can also be used to compute the global false discovery rate (FDR) for each probability used as a threshold to filter the data. In practice, the peptide level data is taken as input to the next level analysis, in which the probabilities are recomputed and the FDR control is carried out at the protein level^{30, 31}.

Adaptive discriminant function

The outline of the adaptive framework is shown in Fig. 1. The analysis starts with the usual routine of applying the EM algorithm for mixture model to compute peptide probabilities using the discriminant search scores based on the coefficients determined using the original ISB 18 protein mix dataset. The computed probabilities are then used to create a new training dataset on-the-fly as follows. In the unsupervised mixture modeling framework, peptide assignments to spectra whose probabilities are greater than 0.9 or less than 0.01 are selected to build up a new training dataset. The selected low probability peptides are assigned class labels 0 (incorrect). For each high probability peptide (> 0.9), its class label is taken to be 1 (correct) or 0 by sampling with a frequency equal to its probability. In addition, peptide assignments with very low probability (typically 0.0001) are excluded from the new training datasets since those are usually low quality spectra and may distort the discriminant function¹¹. The same procedure is applied in the case of the semi-supervised framework, except the negative set is simply created from matches to decoy peptides.

The class label sampling procedure described above is repeated multiple times (10–50 times) for the high probability spectra, to get varied training datasets (same peptide identifications but different class labels). The updated set of discriminant function coefficients is then derived as an average over the discriminant function coefficients obtained for each sampled training dataset. The updated coefficients are used to calculate the new discriminant score for each peptide assignment, which in turn is used to repeat the mixture modeling part and update the probability. The entire procedure is iterated until convergence, i.e. the discriminant function coefficients become stable and no significant change is observed for peptide probabilities.

Top 10 re-ranking

The standard approach considers the top scoring peptide assignment for each spectrum only (ranked based on Xcorr score in the case of SEQUEST). The modeling can be extended to include the runner-up peptides, e.g. to consider top 10 best scoring peptides, with each peptide match treated as an independent identification. The conventional definition is that ΔC_n measures the relative distance of Xcorr between the first and the second-ranked sequences. To extend this to dataset with top 10-ranked sequences, ΔC_n for j th ranked

match (including the top ranked match) is redefined as $\Delta C_{n[j]} = \frac{Xcorr_{[j]} - Xcorr_{[10]}}{Xcorr_{[j]}}$ i.e. as a normalized score obtained from the difference of Xcorr between the j th ranked and the 10th ranked sequences. For those spectra where less than 10 peptide matches are reported by the search tool, the score of the last reported match is used in place of the 10th ranked match. Note that in the adaptive LDA strategy only the top ranked peptides are used to train the new discriminant function since including all top 10-ranked peptides is unnecessary for this part.

The discriminant search score is calculated for all top 10 peptides using the adaptive discriminant function as described above. The mixture modeling with EM algorithm is employed using all the top 10 peptides and the posterior probability of correct identification is calculated for each match. Finally, the order of peptide matches is re-ranked for each spectrum based on the order of their posterior probabilities (from high to low). For each spectrum, the new top-ranked peptide (the one with the highest posterior probability) is retained for subsequent analysis.

Spectrum quality-based clustering

After the initial analysis, MS/MS spectra are separated into three clusters according to their spectrum quality. The spectrum quality score is computed using QualScore³². The quality score is measured on a continuous scale, which is then discretized into three bins: below -1, between -1 and 1 and equal or greater than 1, representing low, intermediate, and high quality spectrum cluster, respectively. The EM mixture modeling procedure is then employed for each cluster separately, which means different mixture distributions are assumed for different clusters. In this work, only the spectrum quality score is used to establish the clusters. However, the method is general and other scores or multiple spectrum properties can be used to cluster the spectra.

RESULTS

The performance of the adaptive discriminant function method, inclusion of top 10 peptide assignments per spectrum with re-ranking, and the effect of spectrum clustering were first investigated using the protein mix datasets. Both unsupervised and semi-supervised mixture modeling frameworks were applied. When the semi-supervised framework was used, in order to allow objective comparison between the extended and the original approaches, half of the decoy matches were randomly selected and considered to be unknown. These peptides

were used to evaluate the discriminating power of computed probabilities. The remaining half of the decoy matches were utilized as the known incorrect matches and their class labels were used by the semi-supervised EM algorithm.

Different types of mass spectrometers

The first analysis was performed to investigate the differences between datasets acquired using different types of mass spectrometers. This was done using spectra generated on four different instruments: Thermo LTQ-FT, Thermo LTQ, Agilent XCT Ultra, and Waters/Micromass QTOF Ultima. In all cases, the spectra were searched with SEQUEST under the same set of search conditions: enzyme semi-constrained mode, large mass tolerance (3.0 Da), see Methods for detail.

For each dataset, the analysis started with the same default set of discriminant function coefficient derived using the original 18 protein mix LCQ ion trap dataset (the updated coefficients can be found in Ref.¹²). The adaptive LDA-EM mixture model was applied until convergence, and the updated discriminant function coefficients were compared. The results of the analysis using the unsupervised mixture modeling are shown in Fig 2a, which plots the ratio of the discriminant coefficient for Xcorr' over that of ΔC_n , c_1/c_2 , for each of the four different mass spectrometers. The ratio of the coefficients reflects the relative significance of the two scores for discriminating between correct and incorrect identifications. Increase in the ratio indicates reduced significance of ΔC_n . Note that the significance of the third score, RankSp (not plotted), is far less than that of Xcorr' and ΔC_n , and does not change the interpretation.

The ratio of these two discriminant coefficients did not show a significant variation. This indicates that the relative importance of these two factors in the discriminant function (assuming same samples and search conditions) was quite similar regardless of the instrument type. It should be noted, however, that Xcorr and ΔC_n are highly correlated variables, which makes the coefficients sensitive to the details of the LDA analysis. Thus, a far more informative is the direct comparison of the numbers of correct identifications at various FDRs obtained with and without adaptive re-training. Figure 2(b) plots the receiver-operating characteristic (ROC) - like curves for doubly charged spectra for each of the four datasets. The results for triply charged spectra, and also using the semi-supervised framework, are similar (data not shown). The plots focus on the region of small FDRs (0.03), which is of most practical significance. The ROC plots show that the two approaches performed equally well in identifying the correct peptide assignments regardless of the instrument type. The only minor benefit of the adaptive approach can be seen in the case of the LTQ platform. Overall, these results indicate that, keeping other parameters fixed, MS/MS datasets generated on different mass spectrometers are not substantially different to require different discriminant functions optimized for each instrument type, consistent with previous reports³³.

Different database search conditions

The discriminating power of various database search scores depends on the number of candidate peptides selected (on average) for scoring against each experimental spectrum in the dataset. This number depends not only on the size of the sequence database, but also on the search parameters used to perform the search. Among them, the most important are the mass tolerance, the enzyme digestion constraint, and the number of variable modifications allowed. To investigate the effect of varying search conditions in more detail, the analysis was performed using the LTQ-FT dataset. Five search modes, from the least constrained Unconst-LW (enzyme unconstrained, large mass tolerance) to the most constrained Tryptic-NW (tryptic peptides only, narrow mass tolerance) were used in the analysis.

Figure 3(a) plots the ratio of the discriminant function coefficient of X_{corr} and ΔC_n for the five sets of SEQUEST search results. As a general trend, the ratio increases as the search parameters become more restrictive. In other words, ΔC_n becomes less reliable as the number of candidate peptides in the database scored against each spectrum decreases. It becomes more sensitive to random effects due to small size of the candidate peptide set, and thus less useful for discrimination.

The number of correct identifications obtained for triply charged spectra is plotted against FDR in Fig. 3 (b) for four of the five search conditions. For the most constrained search (Tryptic-NW), using the adaptive LDA-EM method resulted in a substantial gain in the number of identified peptides at a fixed FDR. There was also a smaller but still noticeable gain in the small FDR region (< 0.01) for the second most constrained search option (Semi-NW). For the less constrained search options (Semi-LW and Unconst-LW), the original model (fixed LDA coefficients) performs as well as the adaptive model. A similar trend was observed for doubly charged spectra, although the improvement was less significant (data not shown).

These results imply that the relative importance of different discriminant factors is dependent on the search conditions. Hence, using fixed discriminant function coefficients may lead to sub-optimal discrimination with datasets searched under conditions very different from those used for training the discriminant function. In those situations, the adaptive LDA-EM approach, which learns the discriminant function dynamically from the data itself, can correct the LDA projection from a sub-optimal plane to an optimal one and improve the discrimination. This is further illustrated in Figure 3(c), which shows X_{corr} vs. ΔC_n scatter plot in the case of Semi-LW and Tryptic-NW datasets. The dash line indicates the optimal separation plane as determined by the LDA analysis. A shift in the optimal plane is apparent when going from the less constrained to the more constrained search conditions - the closer the optimal plane is to the vertical line, the less important ΔC_n is for discrimination. Also of note is a much higher variability of ΔC_n in the case of Tryptic-NW dataset, especially for incorrect identifications, reflecting a significant reduction in the number of candidate database peptides per searched MS/MS spectrum.

The analysis presented above investigated the improvements realized by using the adaptive discriminant function approach under different search conditions. A separate question is what database search parameters are optimal in terms of the overall number of identified peptides. Table 1 shows the total number of peptides assigned by SEQUEST as the top hit that are correct identifications. Overall, the Semi-LW search condition provided the best results. Narrowing down the mass tolerance from 3.0 Da to 0.025 Da and restricting the search to tryptic peptides only (Tryptic-NW) resulted in a loss of approximately 18% of correct peptide assignments, from 4314 peptides to 3683, in the case of 3+ spectra. This can be explained by the fact that, for 3+ spectra in this dataset, approximately 8% of all correct assignments were semi-tryptic peptides, and for about 10% of spectra the instrument software incorrectly determined the monoisotopic peptide mass (the mass of the first or second isotope was reported instead). Similar results were observed for 2+ spectra.

Furthermore, Table 1 shows that the ability to separate between correct and incorrect identifications is also higher in the case of less constrained searches. At a fixed FDR of 0.005, even with the adaptive LDA option, in the case of Tryptic-NW dataset it was possible to extract only 3022 correct peptide assignments to 3+ spectra. This corresponds to the sensitivity of 82% (3022 out of 3683 in total). For Semi-LW dataset, the sensitivity of filtering was significantly higher, 96% (and even higher for the unconstrained search, Unconst-LW). As a result, the difference between these two search options (Semi-LW and Tryptic-NW) in terms of the number of peptides identified at a low 0.005 FDR was 27%, i.e.

even bigger than the difference of 18% observed for the total number of correct assignments. This observation may appear counterintuitive at first: opening up the search space (e.g. using higher mass tolerance than necessary for LTQ-FT data) to include more candidate peptides would be expected to increase the number of false positives. However, when the relevant peptide properties (e.g. the mass accuracy ΔM in this example) are used for computing probabilities using the mixture modeling approach of PeptideProphet, opening up the search space can create a net positive effect. In the same narrow vs. large mass tolerance example, the correct peptides are always centered in the region of ΔM close to 0. In contrast, the incorrect peptides, as the mass tolerance increases, become distributed across a wider range of possible ΔM values^{12, 13}. This effectively enhances the proportion of correct vs. incorrect assignments among peptides with small ΔM . The distributions of ΔM values are learned by PeptideProphet from the data and factored in the computed probabilities. The more distinct are the differences between ΔM distributions among correct and incorrect identifications, the more discriminant the computed probabilities are when used to filter the data. A similar trend is true for other peptide properties, most notably the number of tryptic peptides NTT.

Top 10 peptides with re-ranking

The analysis was further extended to investigate the benefits of going beyond the top-scoring (based on Xcorr score) peptide assignment per spectrum using the LTQ-FT dataset searched under two search conditions, Semi-LW and Tryptic-NW. The analysis presented in this section was carried out using the semi-supervised EM mixture modeling. The semi-supervised method is expected to be more robust since after inclusion of top 10 peptides per spectrum incorrect identifications dominate the overall distribution, which makes unsupervised mixture modeling more difficult.

After applying the adaptive LDA-EM algorithm using top-ranked peptides only, the discriminant score was calculated for all top 10 peptides per spectrum (see Methods). On the last iteration, the EM mixture modeling was applied to compute probabilities for all top 10 peptides. Following this, two options were used: (1) retaining all top 10 matches ('top 10'); (2) selecting the new top-ranked match based on the probability ('new top 1')¹⁶. The global FDR plots for these two options, as well as the original approach (labeled 'old top 1') are shown in Fig. 4. Several trends are apparent. First, keeping all top 10 assignments per spectrum through the final stages of the analysis, as compared to selecting the new rank 1 peptide, had a negative net effect of increased FDR for the same number of correct identifications. Second, in the case of Tryptic-NW dataset, there was no improvement due to top 10 re-scoring and using the new rank 1 peptide assignment as compared to the original approach. As discussed in the preceding section, in the case of highly restrictive database searches Xcorr score is far more important for discrimination than ΔCn . Thus, peptide ranking based on the discriminant score S is not significantly different than that based on Xcorr alone. Furthermore, in the case of Tryptic-NW dataset the distributions of the mass accuracy score ΔM among correct and incorrect identifications are not significantly different and even identical for the NTT parameter (since all candidate peptides are fully tryptic). As a result, the probability-based ranking largely follows the ranking by the discriminant score S , and hence the original Xcorr score. In contrast, for the Semi-LW dataset, re-ranking shows a significant improvement, for both doubly and triply charged spectra. The improved ranking by the probability score in the Semi-LW dataset in part reflects the importance of the ΔCn score, but also the use of additional peptide properties, NTT, NMC, and ΔM . As a result, a large number of correct tryptic peptide identifications that were initially "masked" (outscored based on Xcorr score) by one or several incorrect semi-tryptic peptides with high ΔM value became the top ranking peptide assignments for their corresponding spectra based on the computed probability.

Spectrum quality-based clustering

One of the underlying assumptions of the parametric mixture modeling approach of PeptideProphet is that all peptide assignments from the same class (correct or incorrect) in the dataset are drawn from a single distribution in terms of their discriminant search score distribution. In general, this may not be true for several reasons, one of which is that low quality MS/MS spectra may show a distinct pattern of their search scores as compared to high quality spectra¹⁷. To better understand the implications of this assumption in practical setting, MS/MS spectra were clustered into categories based on their spectrum quality (see Methods). The analysis was performed using the LTQ Semi-LW dataset. The observed histograms of the discriminant search score S for peptide assignments to doubly charged spectra within each spectrum quality cluster are shown in Fig. 5(a). As expected from the definition of the spectrum quality, the high quality cluster had a significantly larger fraction of correct assignments compared to the intermediate and low spectrum quality clusters. In addition, the discriminant score distribution within each cluster was different from that observed for all spectra combined. Nevertheless, Figure 5(b) indicates that in this dataset both models, with and without clustering, produced similar results in terms of identifying the correct assignments and there was no gain in performing the EM algorithm separately in each cluster.

Application to a complex human serum dataset

To access the performance of various methods on more complex datasets, the analysis was repeated using data generated on an LTQ instrument from a human serum sample (see Experimental Data). The SEQUEST search was performed using the Semi-LW option. The sequence database used in the search had an equal number of target and decoy sequences. The matches to decoy peptides were taken as incorrect identifications, whereas the number of correct identifications, for each probability-based filtering threshold was estimated as the number of matches to target sequences minus the number of decoy matches. The FDR was similarly estimated as the ratio of the number of decoy matches to the number of matches to target sequences^{6, 26}.

Figure 6 presents a comparison between the results of applying the following three models: (1) fixed discriminant coefficients, no re-ranking ('top 1 fixed coeff EM'); (2) adaptive LDA ('top 1 with Adaptive LDA-EM'); (3) adaptive LDA, using top 10 peptides with selection of the top-ranked peptide based on probability ('new top 1 with Adaptive LDA-EM'); Overall, the last approach performed the best, as expected based on the analysis of the protein mix data. In considering the adaptive LDA vs. using top 10 with re-ranking as separate effects, the adaptive LDA approach was responsible for the improvement in the very low FDR region (less than 0.03). In the higher FDR region it did not increase the number of correct identifications, since the adaptive LDA method increases not the total number of correct identifications but the discriminating power of computed probabilities to separate correct from incorrect identifications. In contrast, using top 10 with re-ranking increases the total number of correct identifications, but at the expense of slightly worse discriminating power. As a result, re-ranking leads to an increase in the number of correctly identified peptides in the higher FDR region. The spectrum quality-based clustering did not produce a significant change in the results (data not shown).

DISCUSSION

PeptideProphet approach represents a combination of both supervised and unsupervised modeling, with the unsupervised part playing a far more pronounced role. The supervised part of PeptideProphet is related to the calculation of the discriminant search score. Multiple scores are combined into a single score using the discriminant function developed using

training data. The rest of the analysis can be described as unsupervised modeling, in which the discriminant score distributions and the distributions of other peptide features (NTT, NMC, dM, etc) among correct and incorrect identifications are learned from each dataset anew using the EM mixture modeling algorithm. Furthermore, it is often sufficient to use a single database search score, e.g. Mascot ion score or the expectation value in X! Tandem. In those cases, the discriminant function involves simple scaling factors only and the entire method can be described as unsupervised EM mixture modeling approach.

An extension of the mixture modeling approach of PeptideProphet was recently described that can incorporate the knowledge about the class label for some of the peptide identifications¹². For example, the searched database may include decoy peptides sequences that cannot be possibly present in any of the expressed proteins in the organism of interest. Assignments of decoy peptides to MS/MS spectra could thus be labeled as incorrect. The EM mixture modeling algorithm can utilize this knowledge, in which case the method can be described as semi-supervised EM mixture modeling³⁴. However, the discriminant search scores were still computed in the supervised manner, i.e. using the fixed discriminant score coefficients.

The adaptive LDA approach presented here was outlined in our earlier work¹¹, and we used the elements of this approach for performing dynamic spectrum quality assessment³². It shares similarities with other recent approaches designed to reduce the dependence on the training data via dynamic learning^{14, 15}. Moving from the opposite direction, i.e. extending a fully supervised approach²⁵ to accommodate the differences observed between different experimental datasets, Kall *et al.* recently presented a semi-supervised support vector machine (SVM) method and a computational tool Percolator¹⁴. In the case of Percolator, semi-supervised learning is applied to the entire SVM classifier containing all features. LDA used here is concerned with the database search scores only, since all other information is added at a second step and modeled in an unsupervised or semi-supervised fashion already. The method of Percolator requires decoy peptides, whereas the use of decoys is optional in the adaptive LDA method. Percolator utilizes a larger number of features and a more sophisticated classification method (SVM), which can result in improved discrimination. On the other hand, the use of the most informative scores only and the simplicity of LDA should make it more robust when applied to smaller datasets, or datasets of lower quality. The analysis presented in Ref.¹⁵ also employed a dynamic training approach, but with a specific focus on phosphorylated peptides.

The robustness of any computational method is an important consideration. In practice, computational and statistical analysis is performed on datasets of different size, generated using different experimental protocols, and using samples of varying complexity. The size of the dataset may range from hundreds of thousands of spectra in the case of large-scale shotgun proteomic profiling experiments to a few thousand spectra in single LC-MS/MS run on an affinity-enriched protein sample. The adaptive LDA method requires that the dataset contained a sufficient number of correctly identified peptides for constructing a new training dataset for re-training of the discriminant function. The analysis presented in this work indicates that the advantage of the adaptive LDA method is most noticeable in the case of highly constrained searches. Thus, a sufficient and more practical approach would be to implement a simplified treatment in which a different (also pre-computed) discriminant function is applied when the average number of candidate peptides per searched MS/MS spectrum falls below a certain threshold. This may also be beneficial in the case of digestion enzymes other than trypsin¹⁴, which was not investigated in this work. Alternatively, the analysis can be carried out using transformed values^{17, 35} in place of the original Xcorr and ΔC_n scores. It remains to be seen, however, if the conversion from the raw score such as

Xcorr to the expectation value³⁵, or other types of proposed score transformations¹⁷, are not subject to the same concern of increased variability observed for ΔC_n .

Inclusion of more than the top ranked peptide assignment for each spectrum brings only a minor improvement in the number of high confidence peptide identifications, less than 10% increase at FDR around 0.05 and even smaller at lower FDR. This increase diminishes even further at the protein level (data not shown). However, in certain datasets the inclusion of more than the top ranked assignment may be useful for reducing the number of false positive protein identifications arising due to high sequence homology³⁶. For example, in the human serum dataset used in this work, a protein plasminogen-related protein B precursor (SWISS-PROT:Q02325, gene symbol PLGLB1) was erroneously identified with a relatively high ProteinProphet probability of 0.94 due to the identification of a peptides CEEDKEFTCR from 4 MS/MS spectra. However, in all 4 cases, the second ranked peptide assignment was CEDEEFTCR, which corresponds to a related protein, plasminogen precursor (SWISS-PROT:Q00747, gene symbol PLG), unambiguously identified by multiple other peptides unique to that protein. Considering that CEEDKEFTCR contains one missed cleavage, the second ranked peptide (correct sequence) becomes the top scoring peptide based on the computed probability, thus preventing PLGLB1 from being reported (inaccurately) as a separate entry by ProteinProphet. A more effective approach, however, would require developing a unified model that combines peptide-level and protein-level analysis. Such a model would involve grouping of peptides into proteins and computing peptide level-probabilities factoring in the protein grouping information such as the number of sibling peptides simultaneously with the top 10 re-ranking procedure.

The question of what search parameters are optimal for deriving the highest number of correct identifications is of great practical importance. It is particularly relevant in the case of high mass accuracy instruments such as LTQ-FT and LTQ-Orbitrap, where performing searches with very narrow mass tolerance seem to be an attractive option³⁷. It also often debated whether the searches should be limited to tryptic peptides, or at least allowing semi-tryptic peptides^{38, 39}. The datasets used in this work provide some interesting insights. As counterintuitive as it may be, opening up the search space, e.g., using higher mass tolerance than necessary can have a net positive effect, provided the auxiliary information (mass accuracy in this example) is accurately modeled and included in computing the peptide probabilities. It should be noted, however, that the observations made based on the analysis of a small number of datasets should not be generalized. For example, datasets generated using control protein mixtures are known to have a much higher proportion of semi-tryptic peptides as compared to more complex samples due to increased ability of the instrument to sequence lower abundance ions that are a product of in-source decay. The ability to more accurately calculate the precursor ion mass would also affect the choice of the optimal search conditions⁴⁰⁻⁴³. Finally, the results may differ depending on the details of the scoring function implemented in a particular database search tool. Most of the previously reported work was limited to the analysis of database search results followed by simple threshold-based filtering⁴⁴⁻⁴⁷. The future work should include a more detailed analysis on the optimality of the database search conditions, especially in case of high mass accuracy data and in conjunction with more sophisticated post-database search statistical data validation options.

CONCLUSIONS

We presented an adaptive LDA-EM algorithm and compared its performance with the fixed discriminant function - EM mixture modeling approach that is currently implemented in PeptideProphet. The improvements were most noticeable in the case of highly restrictive searches (e.g. narrow mass tolerance, tryptic search) where the number of candidate peptides

in the searched database that is compared against each MS/MS spectrum becomes small. In that case, the delta score measure, such as ΔC_n in SEQUEST, becomes sensitive to random fluctuation and less useful for discrimination than what was observed in the dataset used to train the discriminant function of PeptideProphet. On the other hand, the type of the mass instrument was not a factor, with no improvement observed after re-training the discriminant function for each instrument type. Utilizing top 10 peptide matches per MS/MS spectrum followed by probability-based re-ranking resulted in only a moderate improvement, and only in the case of less constrained searches where a large number of correct peptides were “masked” (not first ranked based on the primary search score) but could be recovered with a help of auxiliary discriminant information. Although the analysis in this work considered peptide identifications derived using SEQUEST and PeptideProphet, most of the observations are general in nature. In particular, it is relevant to other computational tools, e.g. ProteinProspector⁴⁸ and Inspect⁴⁹, that combine multiple search scores into a single discriminant score prior to computing peptide probabilities.

Acknowledgments

This work was supported in part by NIH/NCI Grant R01 CA-126239. We thank Damian Fermin, Xia Cao, David Shiteynberg, Eric Deutsch, Henry Lam, and Bryan Prazen for helpful discussions.

REFERENCES

1. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature*. 2003; 422(6928):198–207. [PubMed: 12634793]
2. Steen H, Mann M. The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol*. 2004; 5(9):699–711. [PubMed: 15340378]
3. Eng JK, McCormack AL, Yates JR. An Approach to Correlate Tandem Mass-Spectral Data of Peptides with Amino-Acid-Sequences in a Protein Database. *Journal of the American Society for Mass Spectrometry*. 1994; 5(11):976–989.
4. Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*. 1999; 20(18):3551–3567. [PubMed: 10612281]
5. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*. 2004; 20(9):1466–1467. [PubMed: 14976030]
6. Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nature Methods*. 2007; 4(10):787–797. [PubMed: 17901868]
7. Chalkley RJ, Hansen KC, Baldwin MA. Bioinformatic methods to exploit mass spectrometric data for proteomic applications. *Biological Mass Spectrometry*. 2005; Vol. 402:289–312.
8. Sadygov RG, Cociorva D, Yates JR. Large-scale database searching using tandem mass spectra: Looking up the answer in the back of the book. *Nature Methods*. 2004; 1(3):195–202. [PubMed: 15789030]
9. Carr S, Aebersold R, Baldwin M, Burlingame A, Clauser K, Nesvizhskii A. The need for guidelines in publication of peptide and protein identification data - Working group on publication guidelines for peptide and protein identification data. *Molecular & Cellular Proteomics*. 2004; 3(6):531–533. [PubMed: 15075378]
10. Xie H, Griffin TJ. Trade-Off between High Sensitivity and Increased Potential for False Positive Peptide Sequence Matches Using a Two-Dimensional Linear Ion Trap for Tandem Mass Spectrometry-Based Proteomics. *J. Proteome Res*. 2006
11. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry*. 2002; 74(20):5383–5392. [PubMed: 12403597]
12. Choi H, Nesvizhskii AI. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *Journal of Proteome Research*. 2008; 7(1):254–265. [PubMed: 18159924]

13. Choi H, Ghosh D, Nesvizhskii AI. Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. *Journal of Proteome Research*. 2008; 7(1):286–292. [PubMed: 18078310]
14. Kall L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*. 2007; 4(11):923–925. [PubMed: 17952086]
15. Du X, Yang F, Manes NP, Stenoien DL, Monroe ME, Adkins JN, States DJ, Purvine SO, Camp IIDG, Smith RD. Linear Discriminant Analysis-Based Estimation of the False Discovery Rate for Phosphopeptide Identifications. *Journal of Proteome Research*. 2008; 7(6):2195–2203. [PubMed: 18422353]
16. Searle BC, Turner M, Nesvizhskii AI. Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. *Journal of Proteome Research*. 2008; 7(1):245–253. [PubMed: 18173222]
17. Martinez-Bartolome S, Navarro P, Martin-Maroto F, Lopez-Ferrer D, Ramos-Fernandez A, Villar M, Garcia-Ruiz JP, Vazquez J. Properties of average score distributions of SEQUEST. *Molecular & Cellular Proteomics*. 2008; 7(6):1135–1145. [PubMed: 18303013]
18. Klimek J, Eddes JS, Hohmann L, Jackson J, Peterson A, Letarte S, Gafken PR, Katz JE, Mallick P, Lee H, Schmidt A, Ossola R, Eng JK, Aebersold R, Martin DB. The Standard Protein Mix Database: A Diverse Data Set To Assist in the Production of Improved Peptide and Protein Identification Software Tools. *Journal of Proteome Research*. 2008; 7(1):96–103. [PubMed: 17711323]
19. Whiteaker JR, Zhang HD, Eng JK, Fang RH, Piening BD, Feng LC, Lorentzen TD, Schoenherr RM, Keane JF, Holzman T, Fitzgibbon M, Lin CW, Zhang H, Cooke K, Liu T, Camp DG, Anderson L, Watts J, Smith RD, McIntosh MW, Paulovich AG. Head-to-head comparison of serum fractionation techniques. *Journal of Proteome Research*. 2007; 6(2):828–836. [PubMed: 17269739]
20. Lopez-Ferrer D, Martinez-Bartolome S, Villar M, Campillos M, Martin-Maroto F, Vazquez J. Statistical model for large-scale peptide identification in databases from tandem mass spectra using SEQUEST. *Analytical Chemistry*. 2004; 76(23):6853–6860. [PubMed: 15571333]
21. Razumovskaya J, Olman V, Xu D, Uberbacher EC, VerBerkmoes NC, Hettich RL, Xu Y. A computational method for assessing peptide-identification reliability in tandem mass spectrometry analysis with SEQUEST. *Proteomics*. 2004; 4(4):961–969. [PubMed: 15048978]
22. Resing KA, Meyer-Arendt K, Mendoza AM, Aveline-Wolf LD, Jonscher KR, Pierce KG, Old WM, Cheung HT, Russell S, Wattawa JL, Goehle GR, Knight RD, Ahn NG. Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Analytical Chemistry*. 2004; 76(13):3556–3568. [PubMed: 15228325]
23. Strittmatter EF, Kangas LJ, Petritis K, Mottaz HM, Anderson GA, Shen YF, Jacobs JM, Camp DG, Smith RD. Application of peptide LC retention time information in a discriminant function for peptide identification by tandem mass spectrometry. *Journal of Proteome Research*. 2004; 3(4):760–769. [PubMed: 15359729]
24. Ulintz PJ, Zhu J, Qin ZHS, Andrews PC. Improved classification of mass spectrometry database search results using newer machine learning approaches. *Molecular & Cellular Proteomics*. 2006; 5(3):497–509. [PubMed: 16321970]
25. Anderson DC, Li WQ, Payan DG, Noble WS. A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: Support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *Journal of Proteome Research*. 2003; 2(2):137–146. [PubMed: 12716127]
26. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*. 2007; 4(3):207–214. [PubMed: 17327847]
27. Choi H, Nesvizhskii AI. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *Journal of Proteome Research*. 2008; 7(1):47–50. [PubMed: 18067251]
28. Fitzgibbon M, Li QH, McIntosh M. Modes of inference for evaluating the confidence of peptide identifications. *Journal of Proteome Research*. 2008; 7(1):35–39. [PubMed: 18067248]

29. Kall L, Storey JD, MacCoss MJ, Noble WS. Posterior error probabilities and false discovery rates: Two sides of the same coin. *Journal of Proteome Research*. 2008; 7(1):40–44. [PubMed: 18052118]
30. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. *Analytical Chemistry*. 2003; 75(17):4646–4658. [PubMed: 14632076]
31. Price TS, Lucitt MB, Wu WC, Austin DJ, Pizarro A, Yocum AK, Ian AB, FitzGerald GA, Grosser T. EBP, a program for protein identification using multiple tandem mass spectrometry datasets. *Molecular & Cellular Proteomics*. 2007; 6(3):527–536. [PubMed: 17164401]
32. Nesvizhskii AI, Roos FF, Grossmann J, Vogelzang M, Eddes JS, Grissem W, Baginsky S, Aebersold R. Dynamic Spectrum Quality Assessment and Iterative Computational Analysis of Shotgun Proteomic Data: Toward More Efficient Identification of Post-translational Modifications, Sequence Polymorphisms, and Novel Peptides. *Mol Cell Proteomics*. 2006; 5(4): 652–670. [PubMed: 16352522]
33. Prazen, B.; Nilsson, E.; Pratt, B.; Sadilek, M.; Martin, D.; Klimek, J.; Gemmill, A.; Hohmann, L.; Jackson, J. Instrument Specific Calibration of PeptideProphet; Seattle, Washington. 54th ASMS Conference on Mass Spectrometry & Allied Topics; 2006.
34. Nigam, K.; McCallum, A.; Mitchell, T. Semi-supervised Text Classification Using EM. In: Chapelle, O.; Zien, A.; Scholkopf, B., editors. *Semi-Supervised Learning*. Boston: MIT Press; 2006.
35. Fenyo D, Beavis RC. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Analytical Chemistry*. 2003; 75(4): 768–774. [PubMed: 12622365]
36. Nesvizhskii AI, Aebersold R. Interpretation of shotgun proteomic data - The protein inference problem. *Molecular & Cellular Proteomics*. 2005; 4(10):1419–1440. [PubMed: 16009968]
37. Olsen JV, de Godoy LMF, Li GQ, Macek B, Mortensen P, Pesch R, Makarov A, Lange O, Horning S, Mann M. Parts per million mass accuracy on an orbitrap mass spectrometer via lock mass injection into a C-trap. *Molecular & Cellular Proteomics*. 2005; 4(12):2010–2021. [PubMed: 16249172]
38. Olsen JV, Ong SE, Mann M. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Molecular & Cellular Proteomics*. 2004; 3(6):608–614. [PubMed: 15034119]
39. Picotti P, Aebersold R, Domon B. The implications of proteolytic background for shotgun proteomics. *Molecular & Cellular Proteomics*. 2007; 6(9):1589–1598. [PubMed: 17533221]
40. Hoopmann MR, Finney GL, MacCoss MJ. High-speed data reduction, feature detection and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass Spectrometry. *Analytical Chemistry*. 2007; 79(15):5620–5632. [PubMed: 17580982]
41. Mayampurath AM, Jaitly N, Purvine SO, Monroe ME, Auberry KJ, Adkins JN, Smith RD. DeconMSn: a software tool for accurate parent ion monoisotopic mass determination for tandem mass spectra. *Bioinformatics*. 2008; 24(7):1021–1023. [PubMed: 18304935]
42. Shin B, Jung HJ, Hyung SW, Kim H, Lee D, Lee C, Yu MH, Lee SW. Postexperiment monoisotopic mass filtering and refinement (PE-MMR) of tandem mass spectrometric data increases accuracy of peptide identification in LC/MS/MS. *Molecular & Cellular Proteomics*. 2008; 7(6):1124–1134. [PubMed: 18303012]
43. Zubarev R, Mann M. On the proper use of mass accuracy in proteomics. *Molecular & Cellular Proteomics*. 2007; 6(3):377–381. [PubMed: 17164402]
44. Bakalarski CE, Haas W, Dephoure NE, Gygi SP. The effects of mass accuracy, data acquisition speed, and search algorithm choice on peptide identification rates in phosphoproteomics. *Analytical and Bioanalytical Chemistry*. 2007; 389(5):1409–1419. [PubMed: 17874083]
45. Brosch M, Swamy S, Hubbard T, Choudhary J. Comparison of mascot and X!Tandem performance for low and high accuracy mass spectrometry and the development of an adjusted Mascot threshold. *Molecular & Cellular Proteomics*. 2008; 7(5):962–970. [PubMed: 18216375]
46. Haas W, Faherty BK, Gerber SA, Elias JE, Beausoleil SA, Bakalarski CE, Li X, Villen J, Gygi SP. Optimization and use of peptide mass measurement accuracy in shotgun proteomics. *Molecular & Cellular Proteomics*. 2006; 5(7):1326–1337. [PubMed: 16635985]

47. Rudnick PA, Wang Y, Evans E, Lee CS, Balgley BM. Large Scale Analysis of MASCOT Results Using a Mass Accuracy-Based THreshold (MATH) Effectively Improves Data Interpretation. *Journal of Proteome Research*. 2005; 4(4):1353–1360. [PubMed: 16083287]
48. Chalkley RJ, Baker PR, Huang L, Hansen KC, Allen NP, Rexach M, Burlingame AL. Comprehensive analysis of a multidimensional liquid chromatography mass spectrometry dataset acquired on a quadrupole selecting, quadrupole collision cell, time-of-flight mass spectrometer - II. New developments in protein prospector allow for reliable and comprehensive automatic analysis of large datasets. *Molecular & Cellular Proteomics*. 2005; 4(8):1194–1204. [PubMed: 15937296]
49. Tanner S, Shu HJ, Frank A, Wang LC, Zandi E, Mumby M, Pevzner PA, Bafna V. InsPecT: Identification of posttranslational modified peptides from tandem mass spectra. *Analytical Chemistry*. 2005; 77(14):4626–4639. [PubMed: 16013882]

Figure 1. Overview of the LDA-EM procedure

Regular analysis: Experimental MS/MS spectra are assigned peptides using SEQUEST or a similar tool. Multiple database search scores are combined into a single discriminant score S , with the coefficients in the discriminant function determined using a training dataset (18 protein mix data). The distributions of discriminant search score S and auxiliary variables available from the digestion (NTT) or MS¹ steps (mass accuracy dM) are modeled using the mixture model EM algorithm, and each peptide identification is assigned a probability of being correct, p . The list of peptide assignments with computed probabilities is then filtered to achieve a desired FDR, or taken as input in the protein-level analysis. *New adoptive component* (grey boxes): Computed peptide probabilities are used to create a new training

dataset on-the-fly. LDA analysis is repeated using this new training data set. The updated discriminant function coefficients are used to calculate the new discriminant score for each peptide assignment, which in turn is used to repeat the mixture modeling part and update the probabilities. The entire procedure is iterated until convergence.

Figure 2. Analysis of data generated using different types of mass spectrometers
(a) Ratio of the SEQUEST discriminant coefficients $c(X_{corr}')$ and $c(\Delta Cn)$ learned using the adaptive LDA approach for 2+ (solid line) and 3+ (dashed line) MS/MS spectra from four different mass spectrometers, LTQ-FT, LTQ, QTOF and Agilent. MS/MS spectra were searched using Semi-LW search option. **(b)** Number of correct identifications from doubly charged spectra plotted against FDR obtained using fixed discriminant coefficients (black solid line) and using adaptive LDA-EM (red dashed line).

Figure 3. Analysis of LTQ-FT data with spectra searched using different database search parameters

(a) Ratio of the SEQUEST discriminant coefficients $c(X_{corr}')$ and $c(\Delta Cn)$ learned using the adaptive LDA approach for 2+ (solid line) and 3+ (dashed line) MS/MS spectra searched with five different search options. **(b)** Number of correct identifications from triply charged spectra plotted against FDR obtained using fixed discriminant coefficients (black solid line) and using adaptive LDA-EM (red dashed line) for four of the five search options. **(c)** ΔCn plotted against X_{corr}' score for randomly selected 20% data from Tryptic-NW dataset (left panel) and Semi-LW dataset (right panel). Incorrect identifications are shown as black

circles, correct as red triangles, grey dashed line represents the discriminant line based on ΔC_n and X_{corr} scores.

Figure 4. Re-scoring top 10 peptide assignments to each MS/MS spectrum

(a) Number of correct identifications in the LTQ-FT dataset, Semi-LW search option, plotted against FDR obtained using three models: fixed discriminant function, using top scoring (based on Xcorr score) peptide assignment only (black solid line); adaptive LDA-EM retaining all top 10 matches (red dashed line); adaptive LDA-EM with selection of the new top-ranked match based on the probability (green dash-dot line), shown separately for 2+ (left panel) and 3+ (right panel) spectra. (b) Same as (a), Tryptic-NW search option.

Figure 5. Distributions of SEQUEST discriminant search scores for spectra of different quality

(a) Histogram of the discriminant search score S plotted for all 2+ MS/MS spectra in the LTQ dataset, as well as plotted separately for spectra of high (cluster 1), intermediate (cluster 2), and low (cluster 3) quality. Correct identifications are colored in blue. **(b)** Number of correct identifications (2+ spectra) plotted against FDR obtained by applying the adaptive LDA - semi-supervised EM approach on all the spectra combined (black solid line), or separately within each spectrum quality cluster (red dashed line).

Figure 6. Application to complex protein samples

Estimated number of correct identifications in the human serum dataset, Semi-LW option, plotted against FDR obtained using three models: fixed discriminant coefficients, no re-ranking (solid black line); adaptive LDA (red dashed line); adaptive LDA, using top 10 peptides with selection of the top-ranked peptide based on probability (green dash dot line). Left and right panels show the results for spectra of 2+ and 3+ charge state, respectively.

Table 1

Total number of correct peptide assignments to MS/MS spectra of 3+ charge state, and the number of correct assignments and sensitivity at 0.005 FDR, in LTQ-FT protein mix dataset searched with SEQUEST using five different search options.

Search Option	Total Correct	Correct (FDR 0.005)	Sensitivity (FDR 0.005)
Tryptic-NW	3683	3022	0.82
Semi-NW	3967	3452	0.87
Unconst-NW	3842	3546	0.92
Semi-LW	4314	4159	0.96
Unconst-LW	4066	3978	0.98