Journal of proteome research

Article

pubs.acs.org/jpr

Learning Score Function Parameters for Improved Spectrum Identification in Tandem Mass Spectrometry Experiments

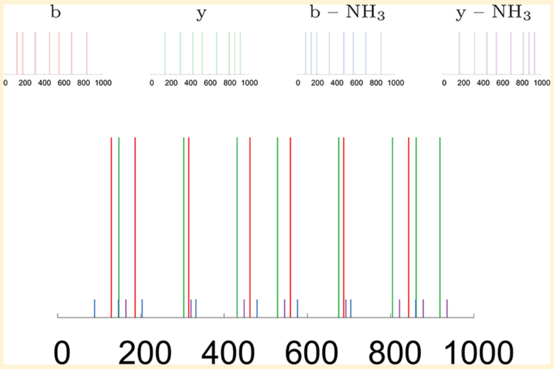
Marina Spivak,[†] Michael S. Bereman,[†] Michael J. MacCoss,[†] and William Stafford Noble[†,‡,*]

[†]Department of Genome Sciences, University of Washington, Seattle, Washington, United States

[‡]Department of Computer Science and Engineering, University of Washington, Seattle, Washington, United States

S *Supporting Information*

ABSTRACT: The identification of proteins from spectra derived from a tandem mass spectrometry experiment involves several challenges: matching each observed spectrum to a peptide sequence, ranking the resulting collection of peptide-spectrum matches, assigning statistical confidence estimates to the matches, and identifying the proteins. The present work addresses algorithms to rank peptide−spectrum matches. Many of these algorithms, such as PeptideProphet, IDPicker, or Q-ranker, follow a similar methodology that includes representing peptide-spectrum matches as feature vectors and using optimization techniques to rank them. We propose a richer and more flexible feature set representation that is based on the parametrization of the SEQUEST XCorr score and that can be used by all of these algorithms. This extended feature set allows a more effective ranking of the peptide-spectrum matches based on the target-decoy strategy, in comparison to a baseline feature set devoid of these XCorr-based features. Ranking using the extended feature set gives 10−40% improvement in the number of distinct peptide identifications relative to a range of q-value thresholds. While this work is inspired by the model of the theoretical spectrum and the similarity measure between spectra used specifically by SEQUEST, the method itself can be applied to the output of any database search. Further, our approach can be trivially extended beyond XCorr to any linear operator that can serve as similarity score between experimental spectra and peptide sequences.



KEYWORDS: *machine learning, spectrum identification, shotgun proteomics*

1. INTRODUCTION

The core problems in the analysis of shotgun proteomics data include mapping each observed spectrum to the sequence of the peptide that generated the spectrum and determining which of these matches are likely to be correct. Methods for matching spectra to peptide sequences (reviewed in ref 1) can be subdivided according to whether they take as input only the observed spectrum—*de novo* methods—or take as input the observed spectrum and a database of peptides, although the distinction between these two types of algorithms is sometimes fuzzy. In this work, we focus on the latter, database search formulation of the peptide identification problem.

In particular, we use as the starting point for our experiments one of the most widely used database search algorithms, SEQUEST.[2] The SEQUEST algorithm generates a theoretical spectrum with fixed peak heights for each candidate peptide in the database—that is, each peptide whose mass lies within a user-specified range of the inferred precursor mass associated with a particular, observed spectrum—and then uses a cross-correlation-based score, XCorr, to measure the similarity between the observed spectra and these idealized theoretical spectra.

Because database search algorithms like SEQUEST will always output the best-scoring match for every observed spectrum, regardless of the quality of the match, a number of algorithms have been developed to assign confidence estimates to the peptide-spectrum matches (PSMs).[3−8] All of these algorithms essentially solve two distinct problems: (1) ranking the matches produced by a search engine in such a way that (ideally) the top of the ranked list is enriched with correct matches, and (2) assigning to each match an estimate of the likelihood that the given match is correct. For the second task, various statistical measures, such as posterior error probabilities, false discovery rate estimates or q-values, have been employed.[9] In this work, we focus on the ranking task, and we use a previously described method[10] to assign statistical confidence estimates.

Although XCorr can be used directly to rank peptide-spectrum matches, combinations of several measures of peptide-spectrum match quality produced by the database search have been shown to yield substantial increases in the numbers of peptides identified with high confidence. Therefore, many state-of-the-art algorithms represent peptide-spectrum matches as feature vectors, composed of a collection of quality measures of the peptide-spectrum matches, as well as

Received: March 9, 2012
Published: August 6, 2012

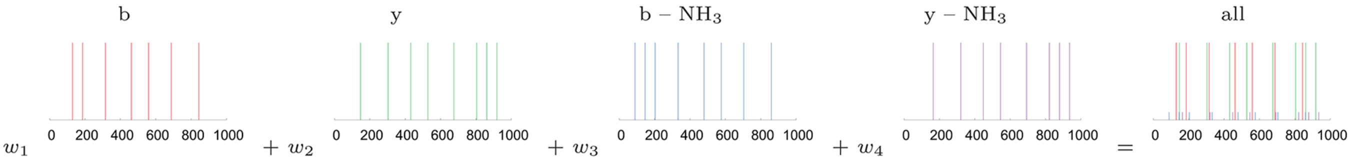


Figure 1. Composition of full theoretical spectrum from weighted sum of subspectra. The figure shows, for the 1^+ charged peptide AGGEFPQRK, theoretical subspectra for b- and y-ions, with and without neutral losses of NH_3. The right-most panel is a sum of these subspectra, with b- and y-ions assigned a height of 1 (corresponding to $w_{1,2} = 1$) and neutral losses assigned a height of 0.1 (corresponding to $w_{3,4} = 0.1$).

characteristics of the candidate peptides and the spectra. For example, for SEQUEST database search, IDPicker[6] uses two indicators of match quality: (1) XCorr (see eq 3), and (2) δC_n, the fractional difference between the current and second best XCorr score. PeptideProphet[3] uses four features: (1) XCorr, (2) δC_n, (3) SpRank, the rank of the current PSM in the list that is sorted based on a preliminary score function, and (4) the absolute value of the difference between the mass of the peptide and the measured mass of the precursor ion. Percolator[7] and Q-ranker[8] further expand this feature set to include charge states, indicators of tryptic termini and other match quality measures (see Supplementary Table 1, Supporting Information).

All of the match verification algorithms mentioned above represent the score of a peptide-spectrum match $\mathbf{x}$ by some family of parametrized functions $f(\mathbf{x})$ and use optimization techniques to find parameters of f based on empirical data. Percolator, Q-ranker and PeptideProphet use discriminant function analysis to determine the parameters of the score function f, while IDPicker uses Monte Carlo simulation to find the optimal parameter values.

The scores of the peptide-spectrum matches based on the values of the function f are then used to estimate statistical measures of the confidence of each match. Thus, Percolator, Q-ranker and IDpicker rank the matches based on their scores and then assign false discovery rates to the matches as an estimate of their correctness based on the ranking. PeptideProphet uses the EM algorithm to assign posterior probabilities that the matches are accurate. At this stage of the analysis, IDPicker and PeptideProphet make use of additional information about the matches—such as enzyme specificity, missed cleavage sites and charge states—that Percolator and Q-Ranker incorporate directly into the input to the peptide-spectrum match score function. PeptideProphet models the joint distribution of the peptide-spectrum match scores, numbers of tryptic termini and numbers of missed cleavages during the EM analysis. IDPicker divides the matches into separate groups based on their charge states and numbers of tryptic termini and then draws FDR thresholds separately for each group.

In this work, we focus on learning the peptide-spectrum match score function $f(\mathbf{x})$. We aim to incorporate elements of the theoretical spectrum generation used during the preceding database search into the learning process. We therefore parametrize the model used by the search engine, and we use machine learning techniques to adjust these parameters during the match evaluation step of the analysis. The main idea in this work is to exploit the linearity of the cross-correlation function as well as the XCorr score in order to parametrize the score function $f(\mathbf{x})$ in terms of the peak heights of the theoretical spectrum. The optimal values of these parameters can then be determined based on empirical data using optimization techniques. The aim is to produce an optimal solution to the target-decoy ranking task. The resulting ranking can then be used for estimations of the statistical confidence of the peptide-spectrum matches.

The advantage of our approach is that we use global information about the whole collection of spectra in the data set to adjust the peak height assumptions made during the database search. Accordingly, the peak height parameters optimal for ranking are estimated in the context of other features of the entire set of peptide-spectrum matches. The information about these global characteristics is not available to the search engine, because the whole peptide-spectrum match collection does not yet exist during the database search.

We demonstrate that the parametrization we propose leads to improved performance in comparison to a baseline feature set that uses fixed peak heights to compute XCorr. In particular, the extended feature set yields 10−40% improvement in the number of peptide identifications over a range of q-value thresholds on all of the data sets examined in this paper. We show that these results are due to the use of the more flexible feature set and are supported by variations in the models that we optimized.

2. PARAMETRIZATION OF THE CROSS-CORRELATION FUNCTION IN TERMS OF THE PEAK HEIGHTS OF THE THEORETICAL SPECTRUM

2.1. SEQUEST Search

Before presenting our parametrized score function, we describe the preprocessing of the observed spectrum, the model of the theoretical spectrum and the XCorr score used by SEQUEST.[2] Before the analysis, each observed spectrum is divided into 10 equal mass-to-charge regions, and each region is linearly normalized by the highest intensity peak contained in it. The theoretical spectrum is represented by ~1 Da bins ranging from 0 to the maximum possible mass-to-charge of peptide fragments in the data set. To create a theoretical spectrum from a peptide amino acid sequence, the algorithm identifies all prefix and suffix ions (b-ions and y-ions), generates six peaks for each ion and distributes these peaks into the bin closest to their mass-to-charge ratio. The six peaks correspond to

1 Primary peaks, with an m/z value based on the sum of the masses of the amino acids in the corresponding b- or y-ion,
2 Flanking peaks, occupying the 1-Th bins on either side of the primary peak, and
3 Three neutral loss peaks corresponding to loss of water (18.0153 Da), ammonia (17.03056 Da) or carbon monoxide (28.0101 Da). The carbon monoxide loss, which is equivalent to an a-ion, is included only for b-ions.

SEQUEST assigns the primary peak a height of 50, flanking peaks heights of 25 and neutral loss peaks heights of 10. All of the peak heights are fixed throughout the search (Figure 1).

The search procedure then computes a similarity measure between each experimental spectrum **s** and theoretical spectrum **t** given by the XCorr score, which measures the extent to which the experimental and theoretical spectra align.[11] The XCorr score is the cross-correlation between spectra with no shift minus the average cross-correlation calculated from a range of shifts:

$$\Xi(\mathbf{s}, \mathbf{t}) = R_0 - \left(\sum_{\tau=-75,\tau\neq 0}^{\tau=+75} R_\tau \right) / 150 \quad (1)$$

where

$$R_\tau = \sum t_i s_{i+\tau} \quad (2)$$

The output is a ranking of candidate peptides according to Ξ.

Eng et al.[12] showed that the SEQUEST XCorr score can be calculated efficiently:

$$\Xi(\mathbf{t}, \mathbf{s}) = \mathbf{t}' \cdot \tilde{\mathbf{s}} \quad (3)$$

where

$$\tilde{s}_i = s_i - \left(\sum_{\tau=-75,\tau\neq 0}^{\tau=75} s_\tau \right) / 150$$

and $\tilde{\mathbf{s}}$ are computed once for each observed spectrum **s**. The formulation of the XCorr given by eq 3 indicates clearly that the function is linear in the theoretical spectrum.

2.2. Parametrization

One of the stronger assumptions made by SEQUEST is the assignment of fixed heights to various types of peaks in the theoretical spectra. Here we address this assumption by parametrizing the theoretical spectrum model and adjusting the parameters based on empirical data.

We start with decisions about which ion types in a spectrum will have peak heights that can be modeled as relatively independent of the peak heights of the other ions types. Since all the peaks in a spectrum arise from a single peptide sample, their heights must have some mutual dependence. However, it has been observed that different ion types can be characterized by their specific ranges of peak heights that remain consistent across multiple spectra in the experiment.[13] This observation implies that there are physical and chemical factors that introduce consistent biases in the heights of the peaks of different ion types, making these heights somewhat independent of the original peptide concentration and of each other. Indeed, extensive studies have used large collections of spectra to elucidate these factors, provide statistical analysis of the peak heights of various ion types and attempt to create computational models of these events (see review in ref 13).

For the purposes of parametrization of the theoretical spectrum, we rely on the observation in these studies that the different ion types can be characterized by the height ranges of their peaks. We assume that the following ion peaks will have characteristic heights: b-ion; y-ion; NH_3 loss from b-ion; H_2O loss from b-ion; NH_3 loss from y-ion; H_2O loss from y-ion; flanking peaks; CO loss from b-ion. However, we stress that the parametrization presented in this paper can be trivially extended to any other subdivision of the theoretical spectrum into ion types.

Each ion type is represented by a separate theoretical "subspectrum", containing peaks with m/z values corresponding to a *single* ion type and with *unit* intensities (Figure 1). A full theoretical spectrum for any peptide sequence can be represented as a weighted sum of the N subspectra $\mathbf{t}_1 \dots \mathbf{t}_N$ corresponding to each separate ion type:

$$\mathrm{t} = \sum_{i=1}^{N} w_i \mathrm{t}_i \quad (4)$$

Because the XCorr score given by eq 3 is linear in the theoretical spectrum, it can be written as

$$\Xi = \sum_{i=1}^{N} w_i P_i \quad (5)$$

where

$$P_i = \mathbf{t}'_i \tilde{\mathbf{s}} \quad (6)$$

On the basis of this parametrization, any feature set representation of the peptide-spectrum match can be augmented to contain several features P_i, representing the sum of shifted cross-correlations of the observed spectrum with the theoretical subspectrum for each individual ion type. In particular, all of the postprocessing algorithms mentioned in the Introduction that use the XCorr score as one of the features—PeptideProphet,[3] IDPicker,[6] and Q-ranker[8]—can trivially augment their feature sets in this fashion. Similarly, feature sets that do not contain the cross-correlation or the XCorr score can be extended to contain these new features.

In this work, we use the peptide-spectrum match feature representation previously described in references 7 and 8, as the base feature set, which we employ as a baseline for all the experiments in the paper. We then modify this feature set by adding the XCorr score of the products $P_1 \dots P_8$ given by eq 5 to the base feature set, while leaving all the of the other components intact. The baseline and extended feature sets are described in Supplementary Table 1(Supporting Information). Note that we do not remove the XCorr score from the base feature set when we add the subspectrum features, because our discriminative model is capable of making use of partially redundant features. We compare the performance of these two feature representations of peptide-spectrum matches in optimization techniques based on different score functions $f(\mathbf{x})$.

3. METHODS AND DATA SETS

3.1. Methods

For a detailed description of the optimization problem setup and methods see the Supporting Information. Here we give the essential elements of our approach.

In this paper we use a target-decoy learning strategy,[7,8] and we assign positive labels $y = 1$ to the peptide-spectrum matches containing real peptides and negative labels $y = -1$ to the peptide-spectrum matches containing decoy peptides. We employ a linear model for computing the peptide-spectrum match score function (see eq 1 in the Supporting Information). We then solve a ranking optimization problem which involves determining the parameters of the PSM score function $f(\mathbf{x})$ such that for every pair of target and decoy PSMs, the target scores higher than the decoy (eq 2 in the Supporting Information). Finally, q-values were assigned based on the ranking induced by the peptide-spectrum match scores as previously described[10] (eq 5 in the Supporting Information). We also compared the performance of existing algorithms: PeptideProphet, Percolator and Q-ranker. Q-ranker was

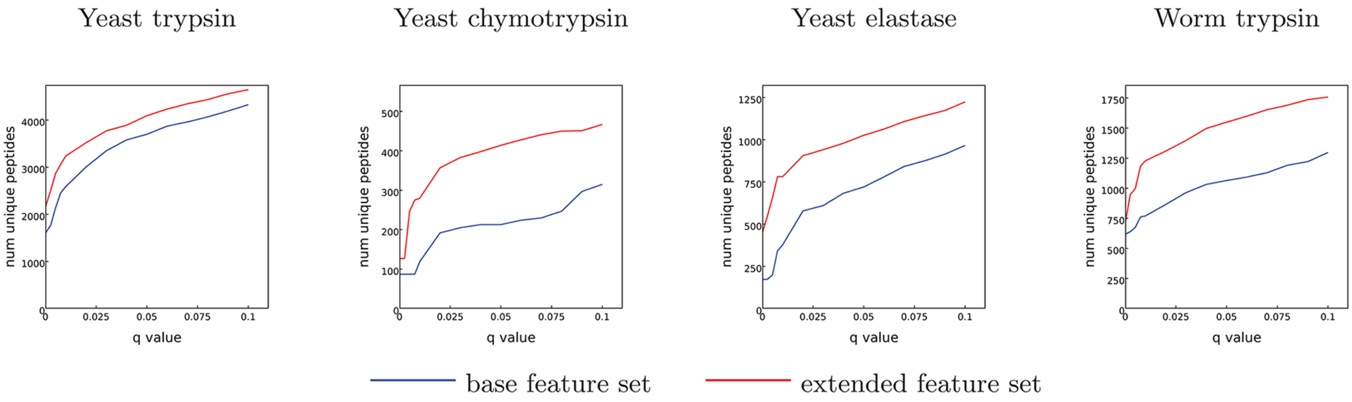


Figure 2. Comparison of base and extended feature sets. Number of unique target peptides identified as a function of *q*-value threshold for the ranking algorithm using base and extended feature sets.

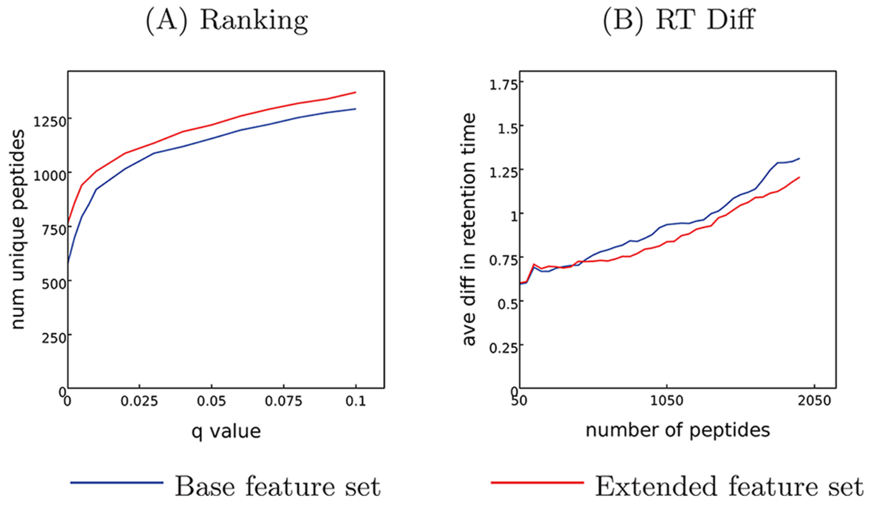


Figure 3. Comparison of base and extended feature sets on six replicate *C. elegans* data sets. (A) Number of unique target peptides identified in two or more replicate data sets as a function of *q*-value threshold for the ranking algorithm using base and extended feature sets. (B) Average of absolute values of retention time differences (in minutes) of peptides identified in two or more replicate data sets as a function of number of peptides at the top of the rank list.

modified to take as input either the base or extended feature set for peptide-spectrum matches.

3.2. Data Sets

We tested the extended feature representation for peptide-spectrum matches on four different sets of data (Supplementary Table 2, Supporting Information). The first set, described in reference 7, consists of three *S. cerevisiae* yeast lysates digested with trypsin, chymotrypsin and elastase as well as a *C. elegans* worm lysate digested with trypsin and analyzed on an Orbitrap mass spectrometer. We refer to these four sets as YT, YC, YE and WT, respectively. The second set, described in reference 14, contains six replicate runs of *C. elegans* lysate digested with trypsin, acquired on a high resolution Orbitrap mass spectrometer. We refer to these data sets as run1—run6. The third set contains eight samples that represent a dilution curve of 48 known proteins synthesized by Sigma (see Supplementary Table 3). These data sets are mixtures (mix1−mix8 in Supplementary Table 2) of the *C. elegans* lysate at equal concentrations and the 48 proteins that are diluted by a factor of 2 in each successive mix, with mix1 having the highest concentration of 840 fmol of the synthesized proteins and mix8 having the lowest concentration of 6 fmol (the concentrations in all the eight data sets are listed in Supplementary Table 1). Finally, to demonstrate the performance of our method on data sets analyzed by different collision-induced dissociation methods, we use two *C. elegans* lysate data sets described in reference 15. The first data set (WHCD in the Supplementray Table 2) was analyzed using a front-end higher energy collision-induced dissociation (fHCD), whereas the second data set (WCID in the Supplementray Table 2) was analyzed using resonance excitation collision-induced dissociation(RE-CID).

For all data sets, peptides were assigned to spectra by using the Crux implementation of the SEQUEST algorithm.[16] The search was performed against a concatenated target-decoy database composed of open reading frames of the corresponding organism and their reversed or randomly shuffled versions, as specified below. All the searches were performed without variable modifications, using a 3.0 Da precursor mass window, and requiring candidates peptides to have at least one enzymatic terminus and no missed cleavages. The top three PSMs for each spectrum were retained for the analysis.

4. RESULTS

4.1. Ranking Results

We first compare the ranking performance using the base and extended feature sets on the three yeast data sets (YT, YC and

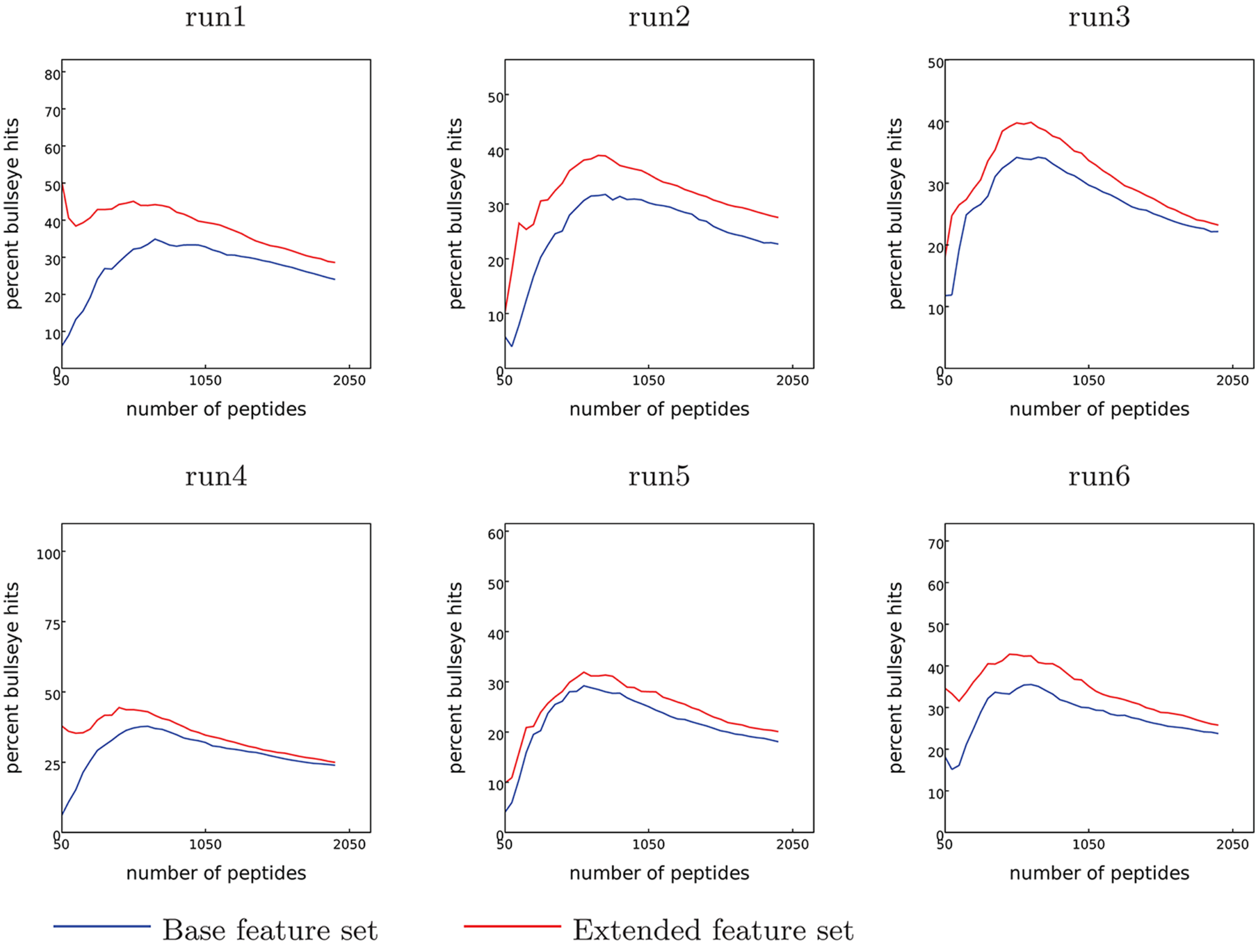


Figure 4. Percent of peptide-spectrum matches that were considered "high quality" by the Bullseye algorithm. Percent of "Bullseye hits" among the peptide-spectrum matches identified using the extended feature set or base feature set as a function of number of peptide-spectrum matches at the top of the ranked list in the six replicate runs.

YE) and one worm data set (WT). We use the linear model to represent the peptide-spectrum match score function, and we optimize a ranking loss function. Figure 2 shows that using the extended feature set leads to superior ranking performance across a wide range of *q*-value thresholds. Supplementary Figure 1 (Supporting Information) computes the percent increase in the number of identified target peptides when using the extended feature set over the base feature set. The percent increase ranges from 10% to 40% over various *q*-value thresholds on all four data sets analyzed.

We further verify that this result is not dependent on the type of the decoys used for the analysis. While Figure 2 and Supplementary Figure 1 (Supporting Information) show analysis on the data sets in which decoys are represented by the reversed peptide sequences, Supplementary Figure 2 shows the same comparison on the data sets which represent decoys as randomly shuffled peptides sequences. Supplementary Figure 3 verifies that the percent improvement due to using the extended feature set again ranges between 10 and 50% on data sets that used random peptide sequences as decoys.

4.2. Ranking on Six Replicate Worm Data Sets

Thus far, all of the comparisons we have presented are contingent upon the target/decoy method for estimating q-values. To control for potential bias in these estimates, we carried out three additional validations that do not depend upon the q-value estimation procedure.

The first of these three methods investigates the reproducibility of our identifications across replicate experiments. We analyze the six replicate *C. elegans* data sets (run1−run6), again using either the base or the extended feature set representations. For the analysis of these data sets, we use the linear model to represent the peptide-spectrum match score function, and we optimize a ranking loss function. Once optimization is performed separately on all the six data sets, the peptides that are identified in two or more data sets are combined into sets for each *q*-value threshold.

Figure 3A shows that, at a range of *q*-value thresholds, the number of replicate peptides identified is higher when the peptide-spectrum matches are represented by the extended feature set. Supplementary Figure 4 (Supporting Information) shows that the percent improvement due to the use of extended feature set ranges from 5 to 30% depending on the *q*-value threshold. The fact that the peptides were identified by at least two out of the six experiments suggests increased confidence of these identifications, in comparison with those obtained by a single experiment.

Table 1. Percent Improvement of the Q-Ranker Algorithm Using the Extended Feature Set Relative to Percolator[a]

	mix1	mix2	mix3	mix4	mix5	mix6	mix7	mix8
concentration (fmol)	870	435	217	109	54	27	13	6
% improvement in total peptide IDs	9.1	10.1	10.3	8.5	9.6	5.7	9.6	8.1
% improvement in known peptide IDs	1.8	2.3	6.6	7.7	8.6	4	0.0	0.0

[a] Percent improvement in the total number of peptides and the number of known peptides identified at $q < 0.01$ by Q-ranker with the extended feature set relative to Percolator with the base feature set.

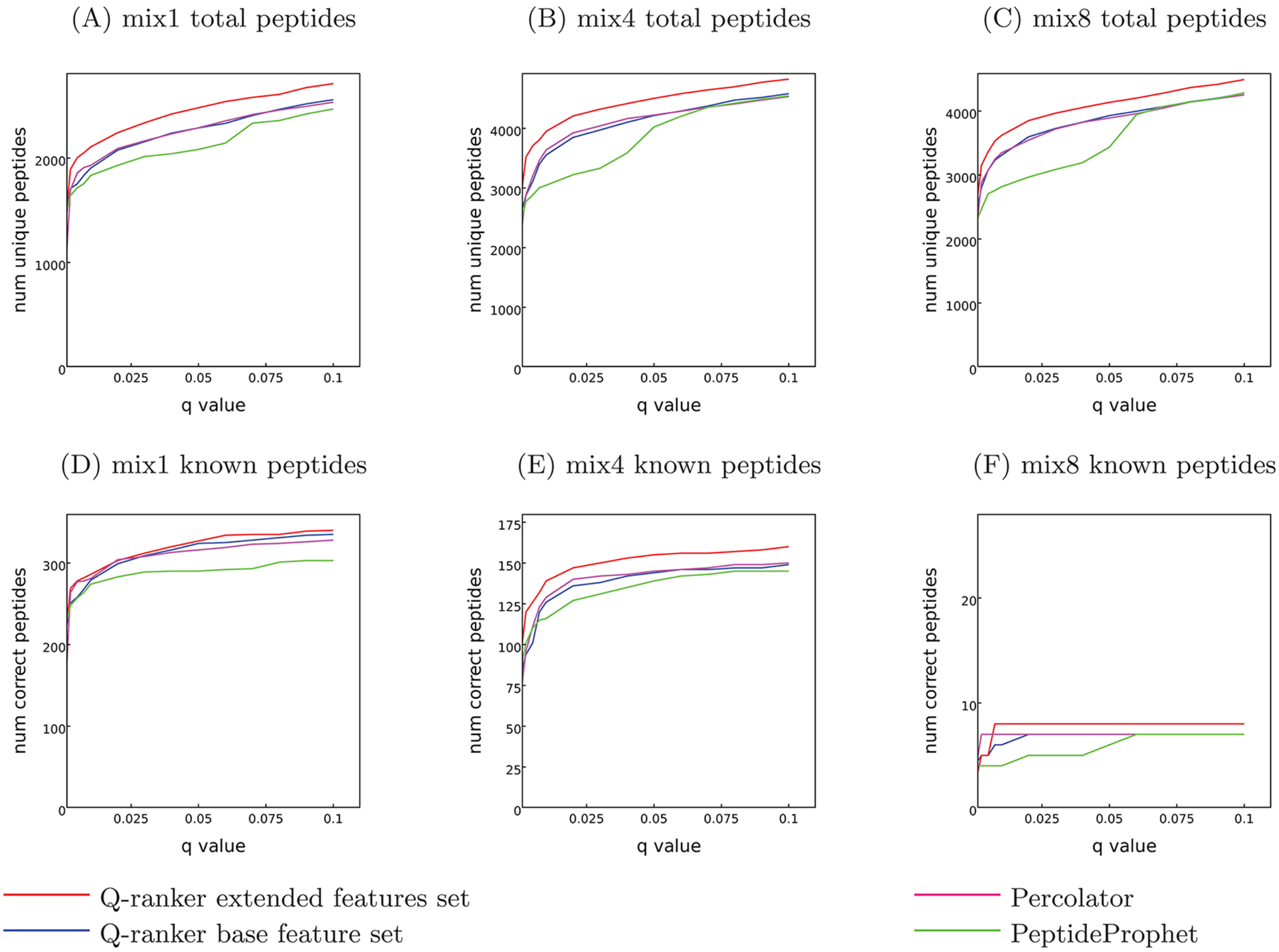


Figure 5. Comparison of PeptideProphet, Percolator and Q-ranker with base and extended feature sets. (A−C) Number of unique target peptides identified as a function of *q*-value threshold. (D−F) Number of *known* target peptides identified as a function of *q*-value threshold.

4.3. Validation by Analyzing Retention Time Differences

The second orthogonal validation compares observed retention times across replicate experiments. The absolute value of the differences in the retention times for the same peptide identified in two or more runs should be small in comparison to the range of the retention times of the peptides during the experiments (from ~0.1 to 99.5 min). For the purposes of validation, we ranked peptides identified using the base and extended feature sets based on their scores and took peptide sets over a range of rank cut-offs to compute the absolute values of their retention time differences across runs. Figure 3B shows that using the extended feature set results in comparable differences in retention times of the peptides identified at a range of rank cut-offs. The peptides identified using the extended feature set do not display unreasonably large retention time differences in comparison with the differences of all the identified peptides.

4.4. Validation by Matching Monoisotopic Mass of Precursor Spectra with Calculated Peptide Mass

Finally, we use an analysis of the precursor spectrum to further validate our method. The precision with which the precursor mass-to-charge of a given fragmentation spectrum can be specified depends upon the precision of the isolation window of the instrument and is often not sufficient to determine accurately the mass-to-charge ratio. The Bullseye software[17] is designed to examine high-resolution Orbitrap data and assign an exact monoisotopic mass to each MS/MS spectrum. This information can serve as a basis for validation of the peptide-spectrum matches. On the one hand, the mass of the peptide in the match can be calculated from its amino acid sequence; on the other hand, the exact monoisotopic mass of the MS/MS spectrum in the match can be assessed using the Bullseye software. Ideally, these masses should be close. In this paper, we considered a peptide-spectrum match to be a “Bullseye hit” if the absolute value of the difference between the calculated peptide mass and the estimated precursor ion mass did not exceed 10 ppm.

We used the Bullseye output to evaluate the peptide-spectrum matches that are identified at various rank cut-offs in the ranked lists produced using either extended or base feature sets. We subjected each of the six *C. elegans* data sets separately to Bullseye analysis and determined the peptide-spectrum matches that contained spectra of high enough quality to receive an assignment of the monoisotopic precursor mass. Further, we calculated the mass of the peptides in these matches based on the amino acid sequences and compared this predicted mass with the estimates of the precursor ion mass determined by the Bullseye.

Figure 4 shows that, in all six replicate *C. elegans* data sets, the analysis with the extended feature representation consistently resulted in a higher percentage of “Bullseye hits” among the identified peptide-spectrum matches at the top of the ranked list, in comparison to the peptide-spectrum matches identified using the base feature set. The percent of hits among the peptides identified at *q*-value less than 0.01 using the extended feature set ranged from 50 to 65%, whereas the percent of hits among the peptides identified at *q*-value 0.01 using the base feature set ranged from 40 to 55% percent.

4.5. Comparison with Existing Algorithms

So far, we have investigated the performance of a simple linear model on the base and extended feature representation of the peptide-spectrum matches. We now compare the performance of the existing state-of-the-art algorithms PeptideProphet,[3] Percolator[7] and Q-ranker.[8] While Percolator and Q-ranker both use the base feature set representation given in the Supplementary Table 1 (Supporting Information), we also modified Q-ranker to take the extended feature set as input, and we compared the resulting performance to that of the other algorithms.

For these experiments, we use eight mixtures (mix1 to mix8 in Supplementary Table 2, Supporting Information) which represent a dilution curve of 48 known synthesized proteins. These mixtures contain equal concentrations of *C. elegans* lysate and successive 2-fold dilutions of the 48 known proteins added to it (concentrations in all the eight data sets are listed in Table 1). The existence of known synthetic proteins in the mixtures allows us to check how many peptides belonging to these proteins are identified, giving a measure of the sensitivity of the analysis by the algorithms compared. Figure 5 presents the results of this comparison on three out of the eight mixtures. Mix1 contains the highest concentration of 870 fmol of the known proteins, mix4 represents one of the intermediate dilutions with concentration 109 fmol, and mix8 contains the lowest concentration of 6 fmol. The results on the other five data sets are found in the Supplementary Figure 5.

Because these data sets consist of *C. elegans* lysate in addition to the 48 known proteins, they contain many more peptides than those derived from the synthetic proteins. Therefore, we first compared the performance of PeptideProphet, Percolator and Q-ranker using base or extended feature sets in terms of the overall number of unique peptides identified by these algorithms at a range of *q*-value thresholds. Figure 5A−C and Supplementary Figure 5A−C and G−H (Supporting Information) show that the Q-ranker algorithm that uses the extended feature set consistently outperforms the other methods in terms of numbers of target peptide identifications. For example, it consistently gives 8% to 10% improvement over Percolator at the *q*-value threshold 0.01 on all the eight data sets (Table 1). Percolator and Q-ranker with the base feature set give comparable results. PeptideProphet identifies fewer target peptides at low *q*-value thresholds, but gives comparable results to the other methods at high *q*-value thresholds.

We then checked how many among all the target peptides identified by each method belonged to the 48 synthetic proteins that were definitely present in the original mixture. Figure 5D−F and Supplementary Figure 5D−F and I−J (Supporting Information) show that for mix1 and mix2, which contain high concentrations of synthetic proteins, all four methods perform equally well on the task of identifying peptides belonging to these proteins. As the concentration of the known proteins decreases (mix3−mix5), using the extended feature set in the Q-ranker algorithm allowed it to identify more peptides belonging to these proteins than the other three algorithms. Q-ranker with the extended feature set showed from 6.6 to 8.6% improvement over Percolator in terms of the number of known peptide identifications at *q*-value threshold 0.01 on these data sets (Table 1). Finally, all algorithms performed equally poorly on the mix6 to mix8, which contained the lowest concentrations of the synthetic proteins.

These results suggest that at sufficiently high concentrations of proteins, the algorithms that do not adjust the peak height parameters during the postprocessing step are still able to successfully identify peptides belonging to these proteins. As the protein concentration decreases, the advantages of the corrections to the original theoretical spectrum model accomplished by using the extended feature set become more significant. Finally, at low concentrations of proteins, the peptides belonging to these proteins become increasingly challenging to identify. However, the poor results on low concentration mixtures may arise due to sampling and detection issues as much as due to informatics. It is likely that at these low concentrations peptides are not detected at all or do not trigger an MS2 scan.

4.6. Examining the Parameter Values Associated with the Theoretical Spectrum Peaks

Because the XCorr score is parametrized in terms of the peak heights of the ions in the theoretical spectrum, we can ask whether the parameters learned during optimization correspond to the peak heights of the experimental spectra. We annotated the peak heights of the spectra contained in the high-confidence peptide-spectrum matches that were identified at *q*-value $\leq$0.01 in the three yeast data sets and one worm data set examined earlier. However, the comparison between these annotated peak heights and peak heights predicted based on the parameters of the XCorr-score revealed very little correspondence (results not shown).

The lack of correspondence between the estimated parameter values and the empirical peak heights can be explained by the fact the we are working in the discriminative rather than generative setting. While generative models are trained specifically to learn accurate peak height ranges characterizing different ion types, discriminative models are trained to give optimal performance on a classification or a ranking task. Because there may be correlations among the peak heights as well as correlation of the peak heights with other

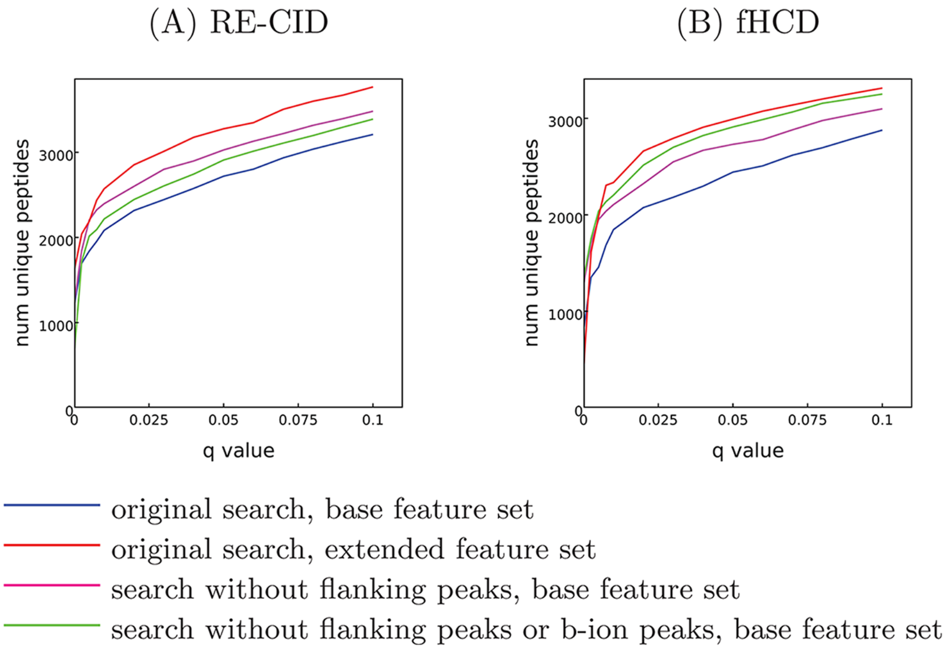


Figure 6. Comparison of RE-CID and fHCD data. Analysis of two *C. elegans* data sets that were generated using either RE-CID and fHCD collision-induced dissociation. The blue and red lines correspond to the database search conducted using theoretical spectrum model with all peaks included, which we call the "original" search. This search was subsequently analyzed using either base or extended feature sets. The cyan line corresponds to the database search that used theoretical spectra without flanking peaks, and subsequent analysis using base feature set. The magenta line corresponds to the database search that used theoretical spectra without flanking peaks or b-ions, and subsequent analysis using base feature set.

features of the peptide-spectrum match representation, the optimal parameters for a discriminative task may not accurately reflect the peak heights in the experimental data.

Despite this caveat, we found that for the majority of the data sets, the parameters associated with the flanking peak heights in the theoretical spectrum were set to zero by the optimization procedure. On the basis of this observation, we hypothesized that the flanking peaks in the theoretical spectrum do not contribute useful information to finding an optimal ranking of the peptides-spectrum matches. To test this hypothesis, we modified the theoretical spectrum model during the search by setting the height of the flanking peaks to zero, while leaving all the other peaks intact. We then analyzed the results of this modified search using the base feature set representation of the peptide-spectrum matches. Finally, we made a comparison of these results with the analysis of the original database search, with all peaks present, using either base or extended feature sets.

Figure 6A shows the results of this comparison on the *C. elegans* WCID data set, analyzed using RE-CID. We find that eliminating the flanking peaks in the theoretical spectrum contributes to more target peptide identifications over a range of *q*-values in comparison with a comparable analysis that includes flanking peaks.

Next, to compare the relative importance of different ion-type features for different collision dissociation methods, we analyzed a *C. elegans* data set (WHCD) that had been produced using fHCD collision dissociation. For this data set, we found that the parameters associated with both flanking peaks and b-ions were consistently set to zero. Therefore, we eliminated both of these ion types from the theoretical spectrum model and repeated the search. Figure 6B shows that elimination of both flanking and b-ions during search contributes to improvement in results when applied to the fHCD *C. elegans* data set. Moreover, the results of this analysis using the base feature set are comparable to the analysis using the extended feature set on the search results using the original theoretical spectrum model. The observation that eliminating b-ions improves the results makes sense due to instability of b-type ions in beam type fragmentation (HCD).[18,19] In contrast, if the b-ions are eliminated from the theoretical spectrum model used during the search of the RE-CID *C. elegans* (WCID) data set, then the number of identifications is decreased (Figure 6A).

Based on these results, we hypothesize that the features associated with b-ions and neutral losses play a more significant role in the analysis of the RE-CID data than fHCD data. Supplementary Table 4 (Supporting Information) shows the results of eliminating different sets of features and documents the percent decrease in the number of the peptide-spectrum match identifications at a q-value threshold of 0.01. We begin by eliminating the original XCorr feature from the extended feature set, because it may compensate for the missing features and interfere with the assessment of their importance. We observe that eliminating XCorr results in a 3% percent decrease in performance. When all the neutral loss features are eliminated, the number of identifications in the RE-CID data is diminished by 9%, whereas the results on the fHCD data stay almost the same, as expected. When both b- and y-ion features are eliminated, then a larger decrease in the number of identifications is observed in both the RE-CID and fHCD data. However, only the RE-CID data is sensitive to the elimination of a single feature corresponding to the b-ion.

This example highlights both the drawbacks and advantages of the approach presented here. On the one hand, the parameters of the XCorr score are not guaranteed to accurately reflect the peak heights of the spectra in the data set, since we are working in the discriminative rather than generative setting. On the other hand, when the discriminant function is given by a linear model, the examination of its parameters can give insights into the usefulness of various features for the discriminative task being addressed.

5. DISCUSSION

While numerous studies have been devoted to elucidation of physical and chemical mechanisms that control systematic peak intensity differences in the spectra produced by peptide ions,[13] the peak heights in a spectrum produced by a given peptide are notoriously hard to predict. Therefore, search engines that rely on generating theoretical spectra to represent peptides in the database are often forced to make strong assumptions about the peak heights in these spectra. In this work, we proposed a method to relax these assumptions during postprocessing of the database search results when the correctness of the matches is evaluated.

This work focused on peptide-spectrum match feature representations that include the SEQUEST XCorr. However, our approach is not dependent on the database search engine, since the method is applied during the postprocessing of the database search results. Our approach only requires that the database search produces a set of measures of the quality of the peptide-spectrum matches that can be used as features to assign scores to these matches. This requirement is fulfilled by most of the widely used search engines, and the feature sets derived from these searches have previously been combined with machine learning techniques for match validation. For example, the Percolator algorithm, originally developed for two feature sets derived from SEQUEST and InsPecT,[20] was later adopted to serve as evaluation tool for the matches produced by Mascot[21] as well. The peptide-spectrum match feature representation of Mascot Percolator replaces the XCorr score with the Mascot score and includes measures of the quality of the spectra and matches similar to the SEQUEST-derived set.[22]

Any such feature set can be augmented with additional cross-correlation-based features, which can be calculated completely independently of the database search itself, as long as the peptide-spectrum matches are available. One of the possible advantages is that the information not used by the search engine due to time and efficiency constraints can be reintroduced during the subsequent analysis. For example, X! Tandem[23] also uses a linear operator as a similarity measure between the observed and theoretical spectrum. It computes a dot product between the observed and theoretical spectrum, multiplied by the factorials of the numbers of b- and y- ions: $N_b!N_y!\Sigma_{i=1}^{n} s_i t_i$. While this model can be factorized itself, it does not contain information about neutral losses of CO, water and ammonia. This extra information can be added in the form of cross-correlation-based features during the evaluation of the peptide-spectrum matches, as described in this paper. The computation of dot product or cross correlation between the observed spectrum and all the theoretical subspectra (as in Figure 1 and eqs 5 and 6) in the postprocessing setting is significantly less time-consuming, since it has to be done only for a single peptide in the peptide-spectrum match, in contrast to multiple peptides during the database search.

However, not every database search algorithm employs linear operators to measure similarity between observed and theoretical spectra. The peptide-spectrum match feature representations derived from such searches can still be augmented by the decompositions described here, since they can be computed completely independently of the search engine. The advantage in this setting is that matching scores from different search engines—that is, the score used by the database search itself and the SEQUEST XCorr-based scores generated during the postprocessing, as described in the paper—can be combined during the statistical evaluation of the quality of the matches.

The approach presented here is flexible and can be easily extended to other representations of the theoretical spectrum and similarity scores between spectra. For example, the theoretical spectrum generation can be extended to more complicated models, which could include separate sets of parameters for each charge state or could take into account characteristics that depend on the precursor mass, the peptide length, the peptide hydrophobicity, etc. Furthermore, this approach could be used by any method that employs linear operators to estimate similarity between two spectra. The XCorr score is a special case of a general linear operator $\mathbf{y}'\mathbf{W}\mathbf{x}$, where $\mathbf{y}$ is a theoretical spectrum, $\mathbf{x}$ is the experimental spectrum and $\mathbf{W}$ is a weight matrix. The matrix $\mathbf{W}$ for the XCorr score has 1s on the diagonal and $-1/150$ for all the other entries. The parametrization of the theoretical spectrum presented in this paper could be used by any method that employs an operator of this form with different entries of the matrix $\mathbf{W}$ to estimate similarity between spectra.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Supplementary methods, tables, and figures. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: william-noble@uw.edu. Tel.: 1 206 221-4973. Fax: 1 206 685-7301.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was funded by NIH awards R01 GM096306 and P41 GM103533.

■ REFERENCES

(1) Nesvizhskii, A. I.; Vitek, O.; Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **2007**, *4* (10), 787−797.

(2) Eng, J. K.; McCormack, A. L.; Yates, J. R., III An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−989.

(3) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identification made by MS/MS and database search. *Anal. Chem.* **2002**, *74*, 5383−5392.

(4) Choi, H.; Nesvizhskii, A. I. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *J. Proteome Res.* **2008**, *7* (1), 254−265.

(5) Ding, Y.; Choi, H.; Nesvizhskii, A. Adaptive discriminant function analysis and reranking of MS/MS database search results for improved peptide identification in shotgun proteomics. *J. Proteome Res.* **2008**, *7* (11), 4878−4889.

(6) Ma, Z.-Q.; Dasari, S.; Chambers, M. C.; Litton, M.; Sobecki, S. M.; Zimmerman, L.; Halvey, P. J.; Schilling, B.; Drake, P. M.; Gibson, B. W.; Tabb, D. L. IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J. Proteome Res.* **2009**, *8* (8), 3872−3881.

(7) Käll, L.; Canterbury, J.; Weston, J.; Noble, W. S.; MacCoss, M. J. A semi-supervised machine learning technique for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4*, 923–25.

(8) Spivak, M.; Weston, J.; Bottou, L.; Käll, L.; Noble, W. S. Improvements to the Percolator algorithm for peptide identification from shotgun proteomics data sets. *J. Proteome Res.* **2009**, *8* (7), 3737–3745.

(9) Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. Posterior error probabilities and false discovery rates: two sides of the same coin. *J. Proteome Res.* **2008**, *7* (1), 40–44.

(10) 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.* **2008**, *7* (1), 29–34.

(11) Eng, J. K.; Searle, B. C.; Clauser, K. R.; Tabb, D. L. A face in the crowd: recognizing peptides through database search. *Mol. Cell. Proteomics* **2011**, *10* (11), R111.009522.

(12) Eng, J. K.; Fischer, B.; Grossman, J.; MacCoss., M. J. A fast SEQUEST cross correlation algorithm. *J. Proteome Res.* **2008**, *7* (10), 4598–4602.

(13) Barton, S. J.; Whittaker, J. C. Review of factors that inflence the abundance of ions produced in a tandem mass spectrometer and statistical methods for discovering these factors. *Mass Spectrom. Rev.* **2009**, *28*, 177–187.

(14) Hoopmann, M. R.; Merrihew, G. E.; von Haller, P. D.; MacCoss, M. J. Post analysis data acquisition for the iterative MS/MS sampling of proteomics mixtures. *J. Proteome Res.* **2009**, *8* (4), 1870–1875.

(15) Bereman, M. S.; Cantabery, J. D.; Egertson, J. D.; Horner, J.; Remes, P. M.; Schwartz, J.; Zabrouskov, V.; MacCoss, M. J. Evaluation of front-end higher energy collision induced dissociation on a bench-top dual pressure linear ion trap mass spectrometer for shotgun proteomics. *Anal. Chem.* **2012**, *84*, 1533–1539.

(16) Park, C. Y.; Klammer, A. A.; Käll, L.; MacCoss, M. P.; Noble., W. S. Rapid and accurate peptide identification from tandem mass spectra. *J. Proteome Res.* **2008**, *7* (7), 3022–3027.

(17) Hsieh, E.; Hoopmann, M.; Maclean, B.; Maccoss, M. Comparison of database search strategies for high precursor mass accuracy MS/MS data. *J. Proteome Res.* **2010**, *9* (2), 1138–1143.

(18) Vachet, R. W.; Ray, K. L.; Glish, G. L. Origin of product ions in the MS/MS spectra of peptides in a quadrupole ion trap. *J. Am. Soc. Mass Spectrom.* **1998**, *9*, 341–344.

(19) Paizs, B.; Suhai, S. Fragmentation pathways of protonated peptides. *Mass Spectrom. Rev.* **2005**, *24*, 508–548.

(20) Tanner, S.; Shu, H.; Frank, A.; Ling-Chi Wang, E.; Zandi, M.; Mumby, P. A.; Pevzner; Bafna, V. InsPecT: Identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.* **2005**, *77*, 4626–4639.

(21) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551–3567.

(22) Brosch, M.; Yu, L.; Hubbard, T.; Choudhary, J. Accurate and sensitive peptide identification with Mascot Percolator. *J. Proteome Res.* **2009**, *8* (6), 3176–3181.

(23) Craig, R.; Beavis, R. C. Tandem: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20*, 1466–1467.