Journal of proteome research

ARTICLE

pubs.acs.org/jpr

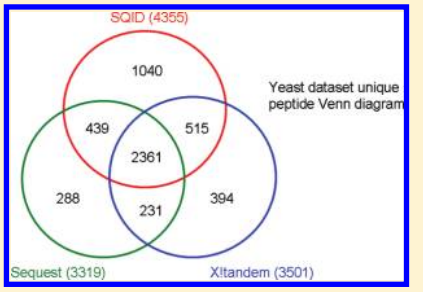
SQID: An Intensity-Incorporated Protein Identification Algorithm for Tandem Mass Spectrometry

Wenzhou Li, Li Ji, Jonathan Goya, Guanhong Tan, and Vicki H. Wysocki*

Department of Chemistry and Biochemistry, University of Arizona, Tucson, Arizona 85721, United States

Supporting Information

ABSTRACT: To interpret LC−MS/MS data in proteomics, most popular protein identification algorithms primarily use predicted fragment m/z values to assign peptide sequences to fragmentation spectra. The intensity information is often undervalued, because it is not as easy to predict and incorporate into algorithms. Nevertheless, the use of intensity to assist peptide identification is an attractive prospect and can potentially improve the confidence of matches and generate more identifications. On the basis of our previously reported study of fragmentation intensity patterns, we developed a protein identification algorithm, SeQuence IDentfication (SQID), that makes use of the coarse intensity from a statistical analysis. The scoring scheme was validated by comparing with Sequest and X!Tandem using three data sets, and the results indicate an improvement in the number of identified peptides, including unique peptides that are not identified by Sequest or X!Tandem. The software and source code are available under the GNU GPL license at http://quiz2.chem.arizona.edu/wysocki/bioinformatics.htm.



KEYWORDS: protein identification algorithm, intensity, tandem mass spectrometry, database search

INTRODUCTION

Tandem mass spectrometry is widely used in proteomic studies because of its ability to identify large numbers of peptides from complex mixtures. In a typical LC−MS/MS experiment, the digested peptides are separated by one or two stages of liquid chromatography and ionized by electrospray ionization. The intact mass of each peptide is measured by mass spectrometry, then the peptide is mass-selected and fragmented to produce MS/MS spectra. These spectra are processed by protein identification algorithms to determine peptide sequences, which, in turn, infer protein sequence. Due to the large number of spectra generated in modern proteomic experiments, protein identification algorithms are a necessity.

Most commonly used algorithms are designed for sequence identification from fragmentation spectra produced by collision induced dissociation (CID), in which peptide precursor ions collide with inert gas molecules and dissociate. CID typically results in fragmentation along the peptide backbone at the amide bonds, producing predominantly N-terminal b and C-terminal y ions. Other ion types, including neutral water and ammonia losses and side chain cleavages, are also possible. Because the masses of the product ions are predictable, the sequence of the original peptide can be reconstructed from the MS/MS spectrum by matching experimental fragment ion masses with theoretical ones. For a long time, m/z has been the main information used by popular algorithms, including Sequest,[1] X!Tandem[2] and Mascot,[3] to assign peptide sequences to fragmentation spectra. The process consists of searching a protein database or translated nucleotide database by m/z for possible peptide candidates, comparing each experimental spectrum with a large number of constructed theoretical spectra or peak lists that correspond to candidate peptide sequences, and assigning a score to each candidate sequence based on the similarity between the theoretical and experimental spectra or on the probability that their match is not random. The strength of the match is finally evaluated according to the top score and the score difference between the top and other candidates.[4]

One limitation of the process described above is that all the major ions of a given series in a theoretical spectrum are assumed to have the same intensity regardless of the properties of the peptide; intensity information contained in an experimental spectrum is essentially abandoned. Though in many cases the m/z information alone is enough to provide reliable identification, intensity can potentially improve the confidence and generate more identifications because it is also highly dependent on the sequence of the peptide and the amino acid residue compositions. Preferential cleavage, for example, is expected at the N-terminus of proline in the presence of a mobile proton or the C-terminus of aspartic acid when no mobile proton is available.[5−7] Nevertheless, intensity is still seldom given much weight in algorithms because of the limited ability to predict and quantify the chemical rules of peptide fragmentation. Various factors, including peptide length, charge state, amino acid content, and charge location, can complicate the process of gas phase peptide dissociation and make the resulting peak intensities difficult to predict and interpret.[8]

Received: September 18, 2010
Published: January 04, 2011

Clearly, the use of intensity to improve peptide identification provides an attractive prospect and efforts have been made by different groups.[9−16] Elias and co-workers, for example, used a probabilistic decision tree—specifically, a treelike feature extracting graph, which requires the members of each branch to have similar properties—to model the probability of observing certain peak intensities in a mass spectrum so as to improve peptide identification.[13] Another algorithm, MASPIC, developed by Narasimhan et al., considered the possibility of random intensity matches as an alternative to using m/z only, based on the assumption that a random match is more likely to correspond to low intensity peaks since these peaks are more common in tandem mass spectra.[14] Zhang reported a kinetic model for prediction of low-energy CID spectra from sequences, assuming that the intensity of a fragment ion is determined by the dissociation pathway and the rate of the dissociation.[9,12] Another intensity model which considers more peptide features and fragmentation rules was developed by Zhou.[16] Intensity is emphasized in these algorithms not because that it is more critical than m/z but because it can provide additional information that can assist with the peptide identification.

The goal of the work presented here is to develop a simple, fast database search algorithm that incorporates rough intensity information to assist peptide identification. In our previously reported study of fragmentation intensity patterns, we introduced a routine to mine a large number of spectra with known sequences for fragment ion intensity based on pairwise amino acid (AA) cleavage patterns, and the relative peak intensity for each AA pair was recorded.[6,7] Because the probability that a data peak of a specific intensity corresponds to any given AA pair cleavage is directly proportional to the probability of that AA pair cleavage resulting in a peak of that intensity, we can evaluate whether the intensity for a certain AA pair in an experimental spectrum is consistent with statistical values. We applied this approach to our SQID algorithm described in this paper. As with other algorithms, the SQID score depends on the presence or absence of ion series peaks at the expected m/z, but is also heavily affected by intensity information to increase the evidence for sequence identification. This is analogous to the manual process of verifying peptide identifications by looking for known fragmentation motifs (e.g., looking for enhanced cleavage at the N-terminus of proline), but with the objectivity of using statistical information gathered in the data-mining process.

METHODS

Algorithm Design

SQID is designed for identification of peptides from ion trap tandem mass spectra in LC−MS/MS experiments but with the ability to extend to spectra acquired using different instruments or dissociation methods (e.g., ETD, ECD) in the future, as long as appropriate training data sets are available. It is written in C language and has been tested in Windows XP and Windows 7 operating systems. The software is available with source code under GNU GPL license at: http://quiz2.chem.arizona.edu/wysocki/bioinformatics.htm. SQID contains a one-time training stage to generate intensity tables that are used in scoring. In the training stage, spectra with known sequence are used to generate the pairwise intensity statistical lookup tables, which quantify the probability of observing a strong peak given a certain amino acid pair. The tables from the training stage are stored in the algorithm and do not need to be regenerated. The scoring process makes use of information from the experimental spectrum and intensity tables to evaluate a match. The algorithm design is described below.

Step 1: Collect Pairwise Cleavage Intensities. The data set used for training contains 138033 unique *D. melanogaster* (version: drosophila-7−14−2008-it) and *S. cerevesiae* (version: yeast-5-04-2009-it) ion trap spectra extracted from the National Institute of Standards and Technology (NIST) Libraries of Peptide Tandem Mass Spectra (http://peptide.nist.gov/).[17] It is a set of spectra with known sequences and consists of singly-, doubly-, and triply charged tryptic peptides ranging from 5 to 56 amino acid residues in length. It contains unmodified peptides as well as peptides with carbamidomethylation of cysteine or oxidation of methionine. Currently we do not treat these modified residues (C+57, M+16) as unique amino acids and their cleavage intensities are combined with those of corresponding unmodified residues (C, M). For each training spectrum, the mass of each expected b and y ion was calculated based on the assigned peptide sequence. Ions outside of the ion trap mass range (high mass cutoff = 2000; low mass cutoff = (precursor m/z)*0.28) were not included (the low and high mass cutoffs can be adjusted as necessary to match the instrument type). The peak intensity of each b and y ion was scaled to the most abundant peak of its own series. The intensity information was sorted by ion type and by the amino acid residue pair cleavage responsible for the fragment ions. Using all training spectra, a histogram was generated containing the relative peak intensities for every expected peak sorted by amino acid pair. When the expected peak was not present, a zero value was included.

Step 2: Calculate Probability of Strong Fragment Ions for Each AA Pair. The relative abundance information for each amino acid pair was separated into three bins: no abundance (intensity = 0), weak (>0−33%) or strong (>33−100%). The ranges defined as weak and strong intensity were empirically determined and the intensity strength for a certain amino acid pair is roughly proportional to the probability of observing a strong peak from that amino acid pair. The probability of having a strong peak (Pr) is defined as the number of strong peaks divided by the total number of expected peaks for the amino acid pair cleavage:

$$\mathrm{Pr} = (\text{number of strong peaks}) / (\text{total number of expected peaks})$$

For instance, the AP pair has a y ion Pr of 0.57, meaning that there is a 57% probability of seeing a cleavage between A and P with a strong y ion peak (>33%). In contrast, the PA pair has a Pr of 0.03, which means that there is only 3% probability of seeing a strong y ion peak for cleavage of the PA pair. In general, these values are in agreement with empirical knowledge and provide a quantitative basis for rough peak intensity prediction given a peptide sequence. Part of the pairwise cleavage intensity probabilities are shown in Table 1. The full table is available in the Supporting Information.

Step 3: Scoring Experimental Spectra. Experimental spectra are assigned peptide sequences by scoring a list of candidate peptide sequences against each spectrum. Each experimental spectrum is modified by eliminating precursor ions, water and ammonia loss products from precursor ions (mass tolerance is the same as fragment tolerance), and isotopes (SQID uses a simple deisotoping algorithm for ion trap data: if the two peaks differ by 1 ± 0.25 and the intensity of the first peak is greater than that of the second one, the second peak is considered to be an isotope peak and removed. The main purpose of deisotoping is to ensure that isotopes of high abundance peaks will not be accidentally selected as top peaks in intensity score calculation). The

Table 1. Pairwise Cleavage Intensity Probability Table for Selected Amino Acid Pairs, Based on Spectra of 138 033 Singly, Doubly, and Triply Charged Unique Peptide Sequences

amino acid pair	Pr of y ion	Pr of b ion
AA	0.23	0.11
AC	0.29	0.10
AD	0.20	0.06
AE	0.17	0.07
AP	0.57	0.28
AQ	0.16	0.09
DN	0.31	0.14
DP	0.70	0.41
PN	0.02	0.01
PP	0.08	0.01
YW	0.35	0.13
YY	0.31	0.11

top 80 of the most abundant peaks from the simplified spectrum are kept for scoring. For each spectrum, a list of candidate peptides (with mass within user-defined tolerance of the precursor mass of the experimental spectrum) is generated from a user-defined FASTA protein database. Each candidate sequence is scored by the following method:

1. Calculate the masses of expected fragment ions from the candidate peptide sequence (same high and low mass cutoffs as in training). In the present work only b and y ions are considered along with H_2O and NH_3 losses from b and y ions. Doubly charged fragments are considered in the circumstance that the precursor ion is triply charged and the mass of the fragment is greater than 900. The value of 900 was empirically determined based on the fact that we seldom see doubly charged ions at $m/z < 450$.
2. Count the number of matched peaks in the experimental spectrum corresponding to the masses of the expected ions for the candidate sequence, within a user defined fragment threshold. If an expected water loss or ammonia loss product is observed, the total number of matched peaks is increased by 0.5. The number of matched fragments is used as a preliminary score and only the top 200 candidates are retained.
3. Count the number of consecutive ion pairs for a match. For instance, if y_5 and y_6 ions are found, it is counted as a consecutive ion pair. Though in many cases consecutive ion pairs increases almost linearly with the number of matched ions, we show later that including them can provide better discrimination than using the number of matched ions alone.
4. For the K most abundant peaks in an experimental spectrum (K depends on the mass of peptide, and equals the integer portion of [2+mass/330]), the Pr of amino acid pairs that result in these peaks are summed and the sum is used as the intensity score: $\sum_{i=1}^{K} \mathrm{Pr}_i$. The intensity score is affected by two factors: how many top peaks are matched and how well the corresponding intensity matches. Because the Pr of amino acid pairs range from 0.01 to 0.72, both factors could play an importance role depending on the sequence.

The final SQID score is calculated as:

$$\text{Score} = (m+n) \times \frac{1 + \sum_{i=1}^{K} \mathrm{Pr}_i}{1 + K \times 0.155} \qquad \text{(eq 1)}$$

where m is the number of matched peaks, n is the number of consecutive ions pairs, Pr is the probability for a certain AA pair to have strong peaks, and K is the number of most intense peaks used to calculate the intensity score. In the scoring function, $(m + n)$ measures the number of matched peaks and numbers of consecutive ion pairs, and increased m and n will increase the confidence of a match; the term $(1 + \Sigma\mathrm{Pr})/(1 + 0.155K)$ measures whether the observed intensity (the numerator) is better than the expected value (the denominator). We expect that the average Pr of the top K peaks is greater than 0.155, the average of all Pr values in the statistical table. A more detailed discussion of the scoring function can be found in a latter section of this report. The specific form of the score function was empirically determined in a trial-and-error manner to reach optimized performance using the Pacific Northwest National Laboratories (PNNL) data set (the first testing data set),[6,7] and then applied to other data sets without any changes. Besides the SQID score, a delta score is used to give further discrimination. Calculation of a delta score in SQID is the same as that in Sequest: the difference of top score and second score was divided by the top score, which shows the percentage difference of the second score to the top score.

The matched peptide sequences and final scores can be reported either as a single tab delimited file or as separate text files. The results can be reported in .OUT format (mimicking Sequest output) for importing into Scaffold[18] to compare with other algorithms, as was done in the present work. Work is in progress to allow future versions of Scaffold to include a separate SQID input.

Performance Test

Data Sets. Three ion trap data sets were used to test the performance of SQID:

1. PNNL data set: contains 28 311 spectra (25% singly charged, 62% doubly charged, and 13% triply charged) from unmodified *Deinococcus radiodurans* and *Shewanella oneidensis* peptides collected on a Thermo LCQ ion trap mass spectrometer.[7,19,20] When these spectra were collected, FT-ICR was used simultaneously for accurate mass measurements. Each LCQ spectrum was then analyzed by the Sequest search engine with *D. radiodurans* and *S. oneidensis* protein databases to assign a sequence. Preliminary identifications of peptides with a minimum cross-correlation score of 1.5 (Xcorr ≥ 1.5) were validated by measurements of Accurate Mass Tags (AMTs) from FT-ICR (mass measurement accuracy < 10 ppm). Though these spectra are of high quality, the error rate of the initial assigned sequences is unclear considering the low Xcorr threshold used. As a result, in current work we use an alternative strategy (see next section) to evaluate the confidence of matches instead of using those initially assigned sequences. The data set is available at http://quiz2.chem.arizona.edu/wysocki/bioinformatics.htm. Spectra were identified against *Deinococcus radiodurans* and *Shewanella oneidensis* database (7984 entries).
2. Eighteen Protein Mixture data set: This data set contains 37 044 spectra collected by the Keller group, Institute for Systems Biology, Seattle, from a mixture of 18 purified proteins using a Thermo Finnigan ESI-ITMS.[21] The data set was collected with 22 LC−MS/MS runs, and only the most abundant peak in each full scan was selected for fragmentation, followed by 3 min of dynamic exclusion. The data set is available at http://regis-web.systemsbiology.net/PublicData

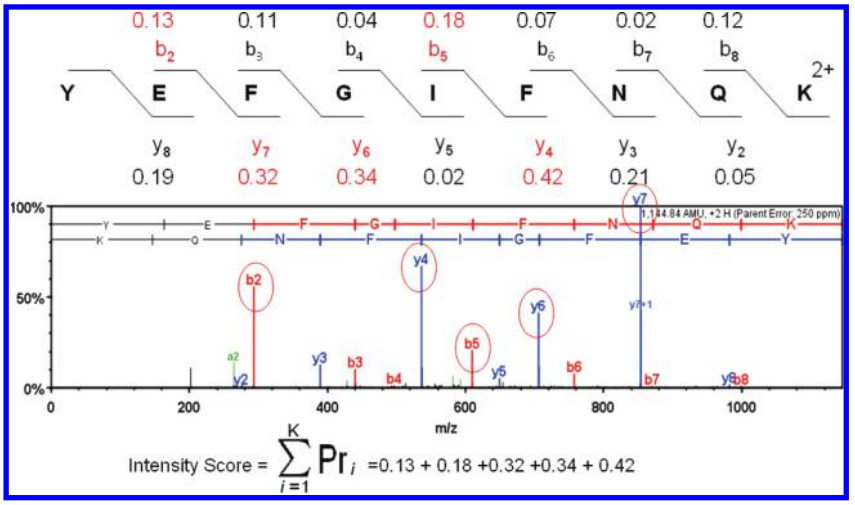


Figure 1. Calculation of intensity score in SQID. The bottom is a labeled experimental spectrum when matching it to the candidate sequence YEFGIFNQK 2+. The most abundant peaks used for the intensity score calculation are circled. The numbers above b ions and below y ions are the probabilities of observing strong peaks with Pr values extracted from the intensity table.

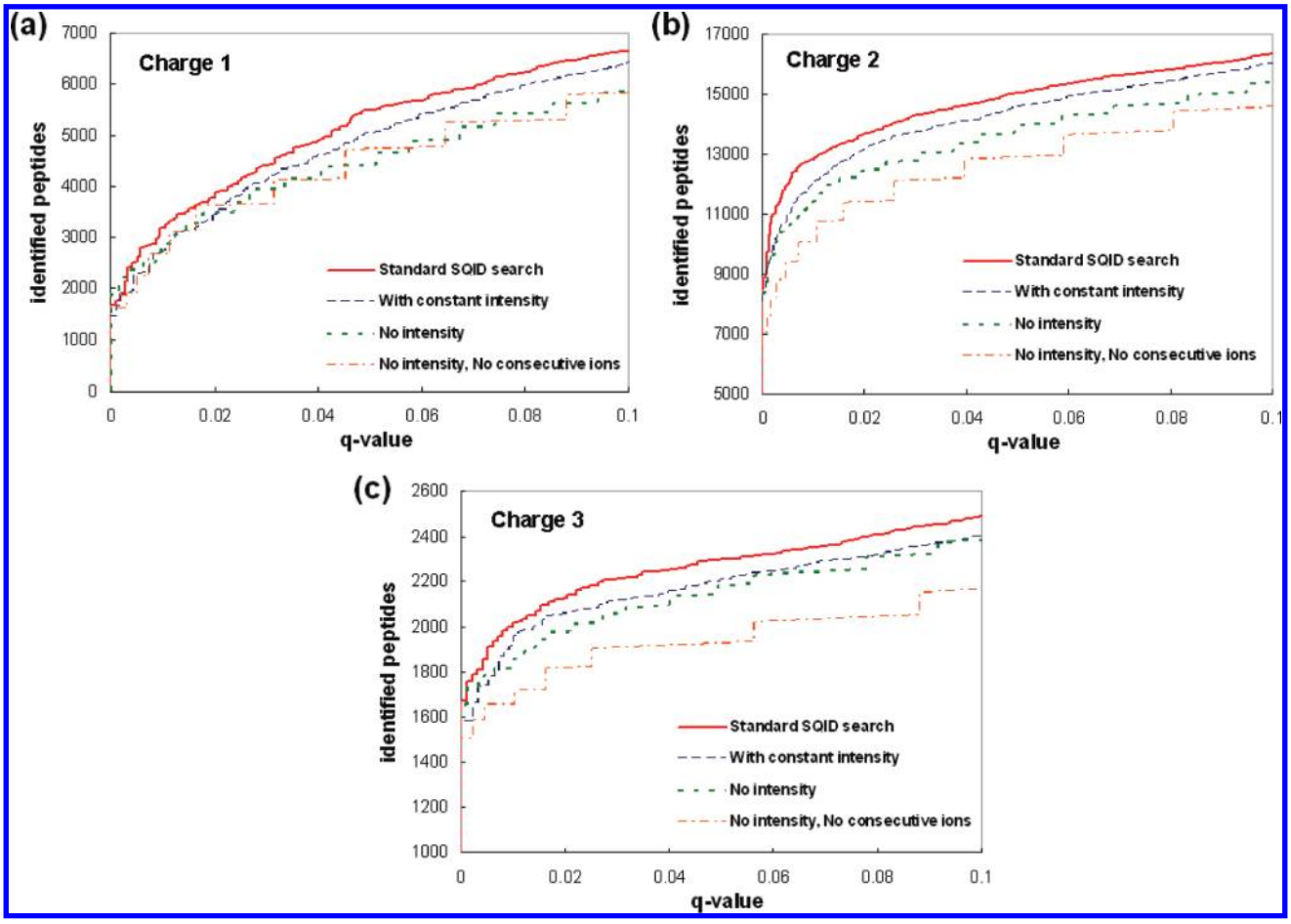


Figure 2. Plot of *q*-value versus number of identified peptides showing the effect of individual components in the SQID score function for (a) singly, (b) doubly, and (c) triply charged peptides. More peptides were identified when adding consecutive ion pairs as well as the intensity related terms to the scoring function.

sets/omics_data set/. Spectra were searched against a reverse version of *Deinococcus radiodurans* and *Shewanella oneidensis* database (7984 entries) plus the 18 protein mixture and common contaminants (trypsin, human keratin, protein standards for MS calibration such as bovine serum albumin and angiotensin, etc).

3. Yeast Data set: This data set of 54 799 spectra from a MudPIT experiment of yeast-extract was collected by the

dx.doi.org/10.1021/pr100959y |*J. Proteome Res.* 2011, 10, 1593–1602

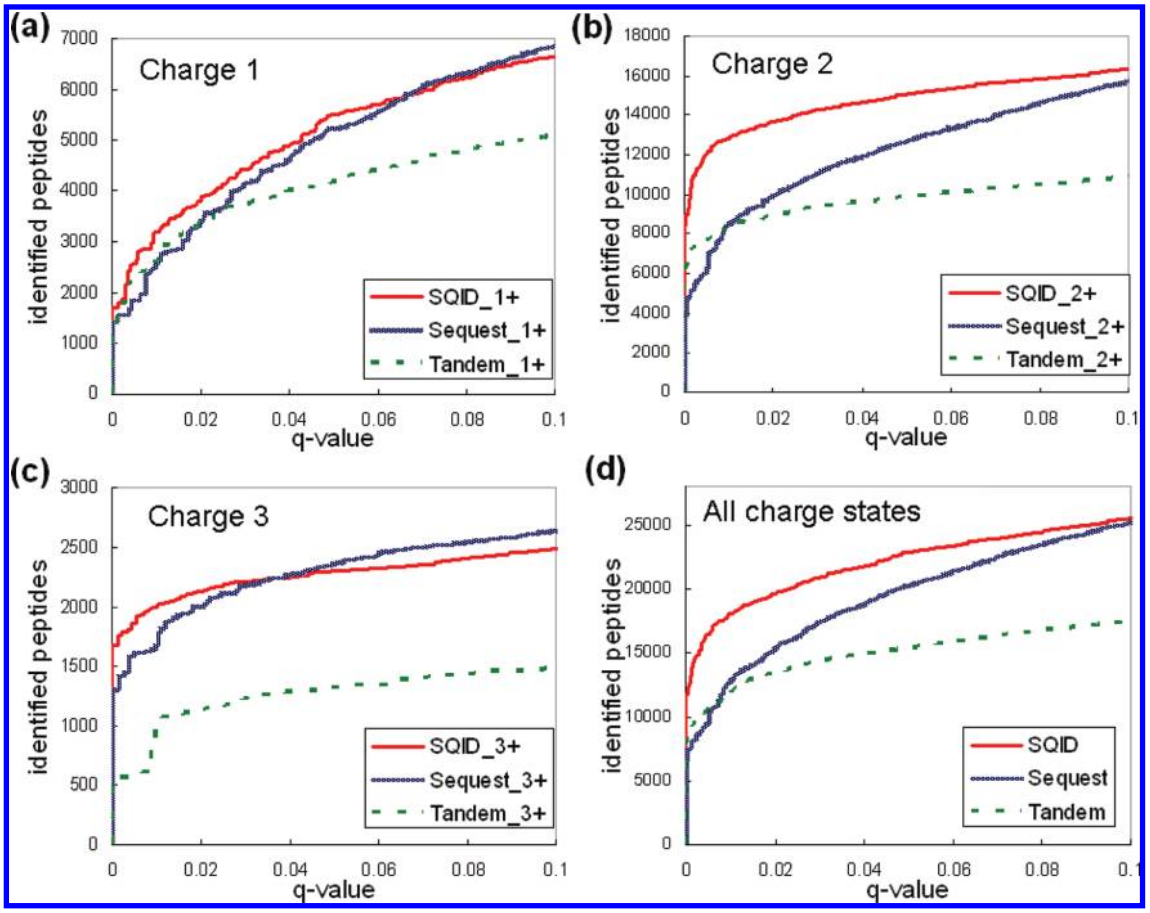


Figure 3. Comparison of SQID, Sequest and X!Tandem by plotting *q*-value (a measure of FDR) versus identified peptide-spectrum matches for the PNNL data set. (a) Singly charged peptides. (b) Doubly charged peptides. (c) Triply charged peptides. (d) Combination of all charge states.

students (group A) of Dr. Andrew Link during the 2006 Cold Spring Harbor Laboratory Proteomics course, on a Thermo LTQ ion trap mass spectrometer.[22] These MudPIT data include six SCX/RP LC separations and the top five most abundant peaks in each full scan were selected for fragmentation. The raw file is available at: http://www.mc.vanderbilt.edu/root/vumc.php?site=msrc/bioinformatics&doc=21164. Spectra were identified against a yeast database (14 590 entries) extracted from the NCBI nonredundant database (ftp.ncbi.nih.gov/blast/db/fasta/). All sequences with "yeast" or "Saccharomyces cerevisiae" in the description line were included.

Search Parameters and False Discovery Rate Determination. The three data sets above were converted to .DTA file format using Bioworks (Version 3.2). Sequest (Version 28, rev.12) and X!Tandem (Version Tornado 2008.02.01.3) were run simultaneously with SQID to evaluate the performance of SQID. Sequest was chosen because it uses a similar scoring that involves no expectation value calculation; X!Tandem was chosen because it is open source and based on expectation values. All algorithms were used with a parent mass tolerance of 1.5 Da and a fragment mass tolerance of 0.5 Da, and a maximum of two missed tryptic cleavage sites. Refinement for X!Tandem was disabled, and the maximum valid E-value for reporting was set to 10 000. PNNL and 18 protein mixture data sets were searched with semitryptic cleavage (tryptic required at one terminus only) and without chemical modifications. The yeast data set was searched with full tryptic cleavage (both termini) and with variable modification of C+57 (carbamidomethylation) and M+16 (oxidation). These modified amino acids are treated as C and M in the SQID intensity score calculation.

For the PNNL data set and the yeast data set, the false discovery rate (FDR) was determined using a target-decoy database search strategy. The database mentioned above was appended with a reverse database using "decoy.pl" program from Matrix Science (http://www.matrixscience.com/help/decoy_help.html#WHAT). At a certain score threshold, the spectra matched to target sequences were labeled "Target" and the ones matched to decoy sequences were labeled "Decoy". The false discovery rate (FDR) was calculated as: FDR = (2 × Decoy)/(Target + Decoy).[23] FDR is further expressed as a *q*-value,[24] which is the minimum FDR threshold at which a given match is considered positive. For easier understanding, the *q*-value can be regarded as a measurement of FDR.

For the 18 protein mixture data set, an identification was assumed to be "True" in the circumstance that the top hit belongs to any of those 18 proteins or common contaminants. At a score threshold which allows "*x*" spectra (among which "*y*" of them are true) to pass, FDR was simply calculated as: FDR = $1 - y/x$.

RESULTS AND DISCUSSION

Calculation of Intensity Score

Intensity information is incorporated into SQID by using statistical intensity tables. Figure 1 is an example of how the intensity score is calculated. For the experimental spectrum with precursor

dx.doi.org/10.1021/pr100959y |J. Proteome Res. 2011, 10, 1593–1602

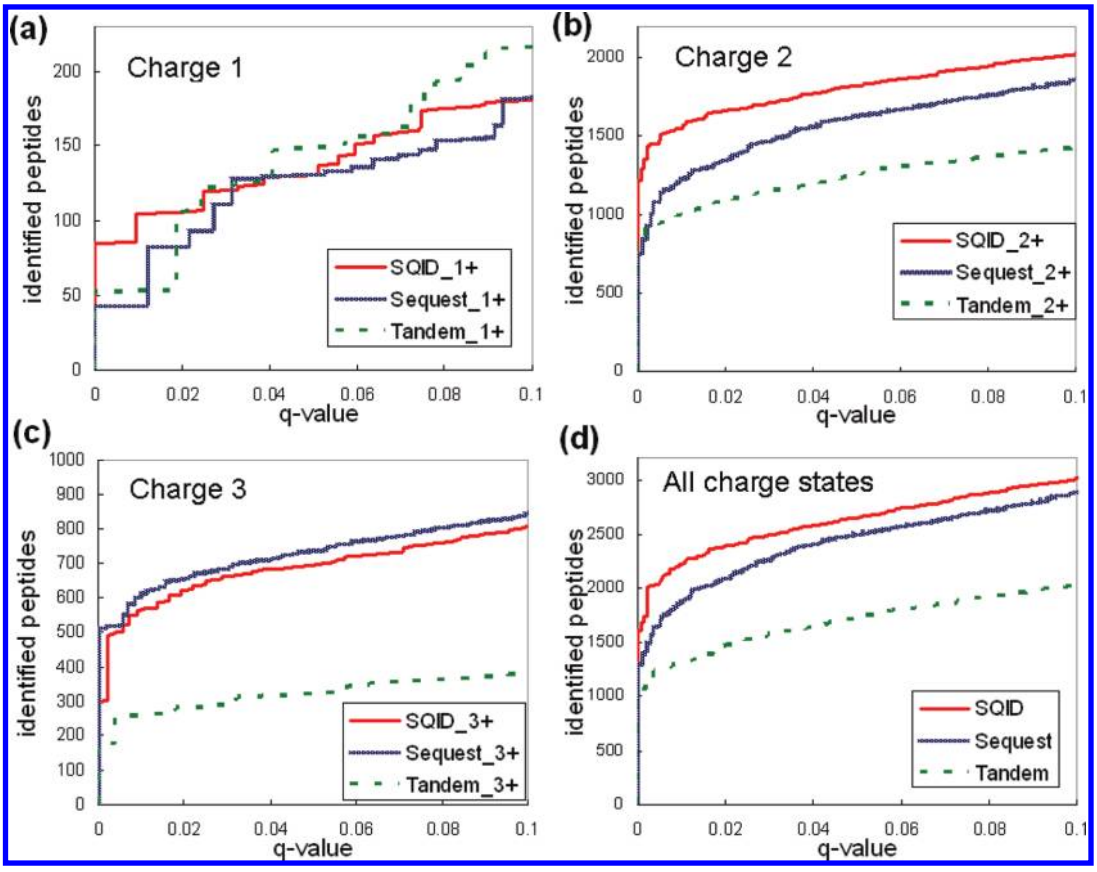


Figure 4. Cmparison of SQID, Sequest and X!Tandem by plotting *q*-value (a measure of FDR) versus identified peptide-spectrum matches for the 18 protein mixture data set. (a) Singly charged peptides. (b) Doubly charged peptides. (c) Triply charged peptides. (d) Combination of all charge states.

MH^+ 1144.8, the top five peaks are used for intensity scoring. Given the candidate sequence YEFGIFNQK^{2+}, SQID will first determine that the top five peaks (integer portion of $[2 + (1144.8/330)]$) are matching to two b ions (b_2, b_5) and three y ions (y_4, y_6, y_7), which correspond to EF and IF pairs for b ions and EF, FG and IF pairs for y ions. By looking up the intensity table, the probabilities to have strong (>33%) peaks (Pr) for each ion pair are 0.13 (EF pair, b_2 ion), 0.18 (IF pair, b_5 ion), 0.32 (EF pair, y_7 ion), 0.34 (FG pair, y_6 ion) and 0.42 (IF pair, y_4 ion). The sum of the above values returns the intensity score. From the graph, it can be clearly seen that the Pr of the top five peaks (shown in red) are among the largest compared with Pr for other peaks (in black), which means that the most abundant peaks in the spectrum are also expected to be statistically strong based on the training set. In general, a higher intensity score indicates that the statistical fragmentation trends are reflected in the match so the confidence of the identification is increased.

Effect of Individual Components in the SQID Score Function

In addition to the number of matched ions used in most algorithms, the SQID score function (eq 1) involves two features to improve peptide identification: consecutive ion series n and intensity $(1+\Sigma Pr)/(1 + 0.155K)$. To evaluate their contributions, four searches were conducted using the PNNL data set: (1) A "Standard SQID search" using eq 1, with *Pr* values adopted from probability table (variable intensity). (2) "With constant intensity" for each ion type, where the Pr value is 0.22 for y ions and 0.09 for b ions. These are the average Pr values for each ion type. (3) "No intensity": both $\sum_{i=1}^{K} Pr_i$ and K equal zero, which completely removes the effect of intensity and ion type. The score function equals $m + n$. (4) "No intensity, no consecutive ion series". The score equals the number of matched ions m. For each search, the results were ranked by the top scores from high to low, then FDR and *q*-values were determined as described earlier. By plotting *q*-value versus the number of peptide hits, Figure 2 showed that more peptides were identified when adding consecutive ion pairs as well as the intensity related terms to the scoring function. From the plots, it should be noticed that by using the number of matched ions alone (m), a significant number of peptides can be identified (orange dot-dashed lines). This illustrates that m/z is powerful information for peptide identification. By adding consecutive ion series (score function is now $m + n$), the performance increases as charge state increases (green dotted lines). This may be explained by the fact that higher charged peptides normally have longer sequences and more theoretical peaks, which will increase the chance of finding consecutive ion pairs. At 0.05 *q*-value cutoff, the performance improved 8% for doubly charged spectra. Adding an intensity term with a constant intensity (blue dashed line) will give a score bonus when a theoretical peak is matched to a high abundance peak, with a higher bonus for y ions and lower bonus for b ions. This step gave an additional 6% (based on "no intensity search") for doubly charged spectra at 0.05 *q*-value cutoff. Finally, the standard SQID score (red solid line), which gives a statistically determined score bonus when a theoretical peak is matched to a high abundance peak, improved the overall performance by another 4% (based on "constant intensity") for doubly charged spectra at 0.05 *q*-value cutoff. The actual performance boost differed for different charge states and *q*-value cutoffs, for example, at 0.005 *q*-value cutoff (2+), "Standard SQID search" outperformed "constant intensity" by 13.4% if the

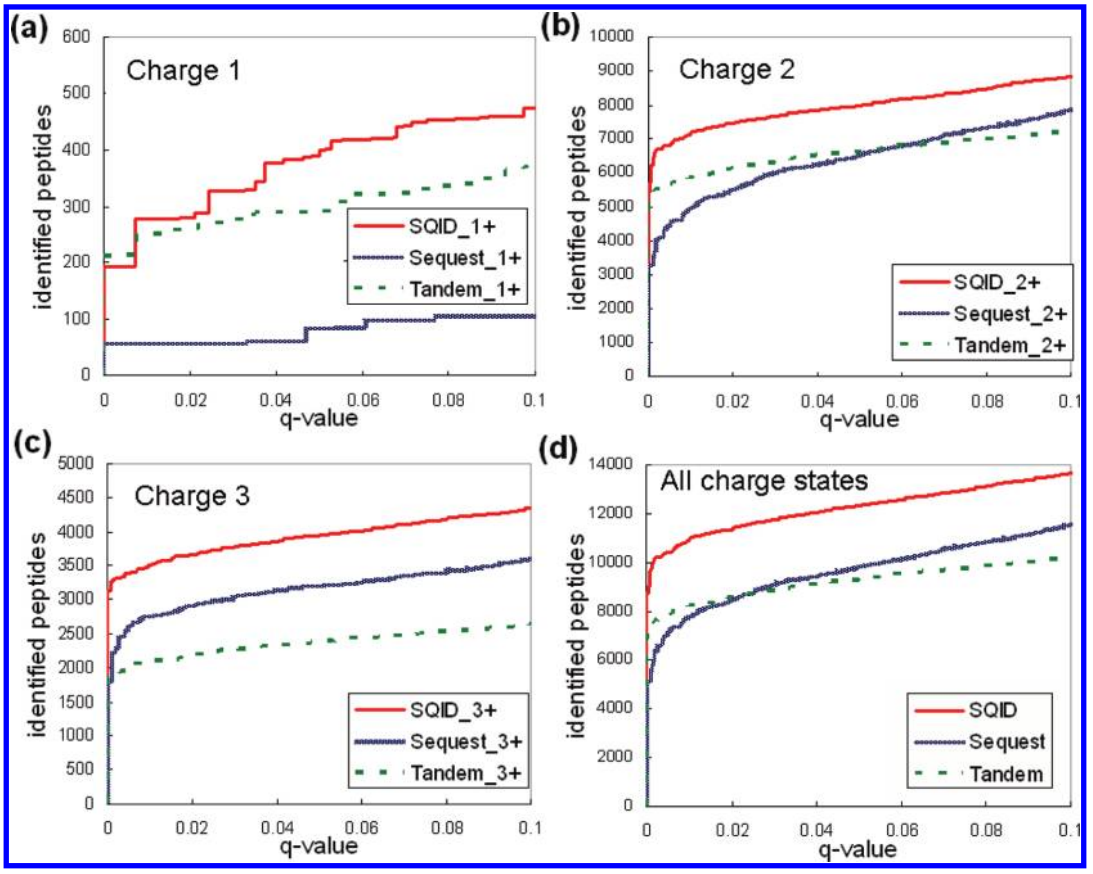


Figure 5. Comparison of SQID, Sequest and X!Tandem by plotting *q*-value (a measure of FDR) versus identified peptide-spectrum matches for the yeast data set. (a) Singly charged peptides. (b) Doubly charged peptides. (c) Triply charged peptides. (d) Combination of all charge states.

Table 2. Unique Peptide Overlap Table for PNNL Data Set at 0.05 *q*-Value Cutoff[a]

	charge			
ID set	1+	2+	3+	total
SQID, Sequest, Tandem overlap	2886	7777	860	11 523
SQID, Sequest overlap	1146	3695	983	5824
SQID, Tandem overlap	680	1198	95	1973
Sequest, Tandem overlap	227	88	89	404
SQID only	625	1889	301	2815
Sequest only	832	730	365	1927
Tandem only	259	479	240	978

[a] Total of 22 135 unique peptides are identified by SQID, compared with 19 678 by Sequest and 14 878 by X!Tandem.

Table 3. Unique Peptide Overlap Table for 18 Protein Mixture Data Set at 0.05 *q*-Value Cutoff[a]

	charge			
ID set	1+	2+	3+	total
SQID, Sequest, Tandem overlap	18	139	47	204
SQID, Sequest overlap	3	24	16	43
SQID, Tandem overlap	4	9	1	14
Sequest, Tandem overlap	2	2	7	11
SQID only	0	12	19	31
Sequest only	3	3	9	15
Tandem only	8	2	2	12

[a] Total of 292 unique peptides are identified by SQID, compared with 273 by Sequest and 241 by X!Tandem.

differences between "No intensity" (green dotted line) and "Standard SQID search" (red solid line) were counted as the contribution of intensity incorporation, normally the value was in the range of 5−15%.

Comparison of Algorithms

A completely objective comparison of algorithms is always difficult because each algorithm uses different spectrum preprocessing methods, different scoring schemes and different score reporting. In the spectrum preprocessing step, Sequest preprocesses the spectrum by keeping the top 200 peaks and separates the spectrum into ten bins for normalization. SQID keeps the top 80 peaks after removing parent related peaks and obvious nonmonoisotopic peaks. X!Tandem simply keeps the 50 most abundant peaks by default. In term of score report, X!Tandem reports "E-value" and a much less important hyperscore, while SQID and Sequest report the main scores as well as delta scores. In this work, the default spectral preprocessing methods were used, and only the main scores, Xcorr, SQID score and E-value were used for a relatively fair comparison. It is important to note that these parameters can potentially affect the search results demonstrated below.

SQID was compared with Sequest and X!Tandem using the PNNL, 18 protein mixture, and yeast data sets. The main scores for each algorithm, SQID score, Xcorr and E-value, were sorted for filtering and *q*-value determination. Figures 3, 4, and 5 compares the

dx.doi.org/10.1021/pr100959y |*J. Proteome Res.* 2011, 10, 1593–1602

search results of SQID, Sequest and Tandem for each data set, at each charge state. Tables 2, 3, and 4 list the unique peptide (no duplicated sequences) overlap table for these data sets at 0.05 *q*-value cutoffs. The performance of SQID varied for different data sets, charge states and *q*-values. For the PNNL data set, it can be seen that SQID yielded similar performance with Sequest for singly and triply charged peptides, but had much more identification for doubly charged peptides, especially at low *q*-value cutoffs. At a *q*-value of 0.05, a total of 22 135 unique peptides were identified by SQID, compared with 19678 by Sequest and 14 878 by X!Tandem (12% and 48% more identification). The 18 protein mixture data set showed a smaller difference between SQID and Sequest at all charge states, but X!Tandem still lagged behind. At 0.05 *q*-value cutoff, 292, 273, and 241 unique peptides were identified by SQID, Sequest and X!Tandem, respectively. For the yeast data set, SQID exhibited strong performances for all charge states in a wide confidence range. At *q*-value cutoff 0.05, the number of unique peptides lead Sequest or X!Tandem by 25% (4355 for SQID, 3319 for Sequest and 3501 for X!Tandem). It was also noted that compared with X!Tandem, Sequest showed a reduced performance for this data set. This may be due to the fact that the spectra are relatively noisy, and Sequest relies primarily on the number of matched ions and keeps more peaks in spectrum preprocessing. For all three data sets, SQID identified a significant number of unique peptides that were not identified by either Sequest or X!Tandem, and the overlap regions between SQID and Sequest or SQID and X!Tandem were normally larger than the region between Sequest and X!Tandem (Tables 2, 3, and 4). SQID also showed a better discrimination power at lower *q*-value cutoffs, which can be seen from the figures.

Table 4. Unique Peptide Overlap Table for Yeast Data Set at 0.05 *q*-Value Cutoff[a]

	charge			
ID set	1+	2+	3+	total
SQID, Sequest, Tandem_overlap	42	1595	724	2361
SQID, Sequest overlap	3	227	209	439
SQID, Tandem overlap	83	349	83	515
Sequest, Tandem overlap	0	137	94	231
SQID unique	57	580	403	1040
Sequest unique	1	166	121	288
Tandem unique	15	246	133	394

[a] Total of 4355 unique peptides are identified by SQID, compared with 3319 by Sequest and 3501 by X!Tandem.

The performance difference between SQID and other algorithms can be attributed to many factors including spectrum quality, spectrum preprocessing, the incorporation of intensity, consecutive ion series, etc., and it is very difficult to individually quantify each of these terms. Here we show as examples two spectra where intensity scoring played important roles. Figure 6a is an example peptide (TKIPAVFK 2+, from one of the 18 proteins) that was identified by SQID but missed by both Sequest and X!Tandem. The spectrum contains two dominant cleavages (IP, KI) and the other fragments are very small. SQID identified this peptide with 8.5 fragments, 5 consecutive ion series, and a very high intensity score 1.44 (for the top 4 peaks). The final SQID score was 20.33, which was 100% confident (*q*-value equals 0 at 20.33). This peptide ranked eighth in Sequest with a Xcorr 1.61; X!Tandem missed this identification, and got a wrong hit with E-value 4.7. In contrast, Figure 6b is a peptide (AAANFFSASCVPCADQSSFPK 2+, from one of the 18 proteins) that was identified by Sequest and X!Tandem but missed by SQID. Though there were a high number of fragment matches, the most abundant peaks essentially can not be matched to any fragments, which gave an intensity score 0. This greatly reduced the final SQID score (the peptide scored 5.35, ranked sixth). Sequest identified this peptide as a top hit with Xcorr 3.73 (100% confident) and X!Tandem identified it with E-value 0.0084 (99% confident).

Further Comparison with Sequest

Because neither SQID nor Sequest are probability based, it is more informative to compare their scores, especially when they

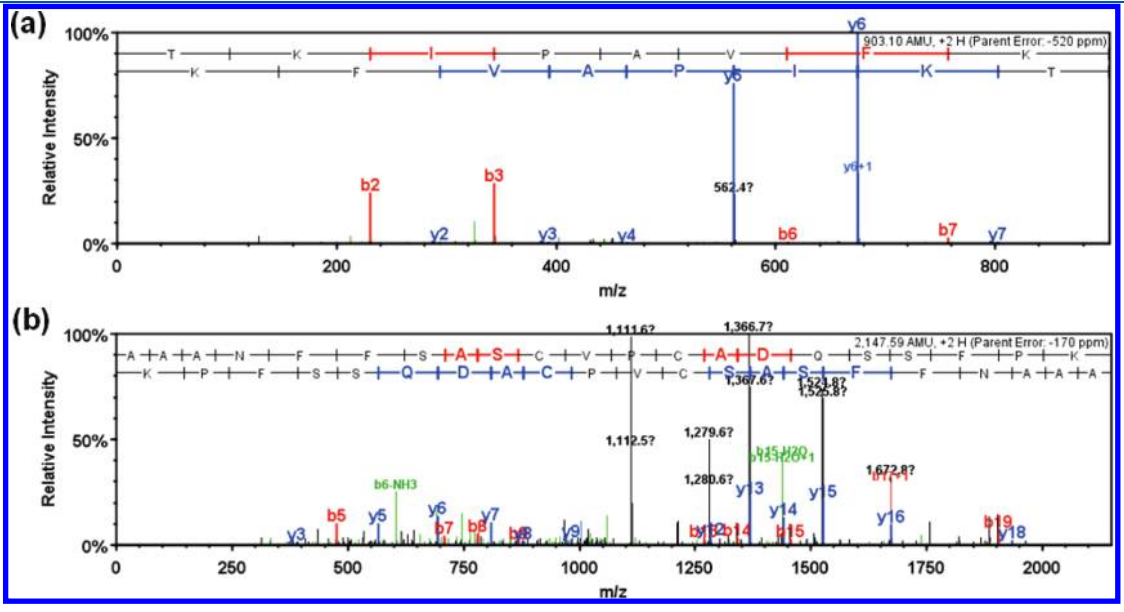


Figure 6. Example spectra that are (a) identified by SQID but missed by Sequest and X!Tandem (TKIPAVFK 2+) and (b) identified by Sequest and X!Tandem but missed by SQID (AAANFFSASCVPCADQSSFPK 2+).

dx.doi.org/10.1021/pr100959y |*J. Proteome Res.* 2011, 10, 1593–1602

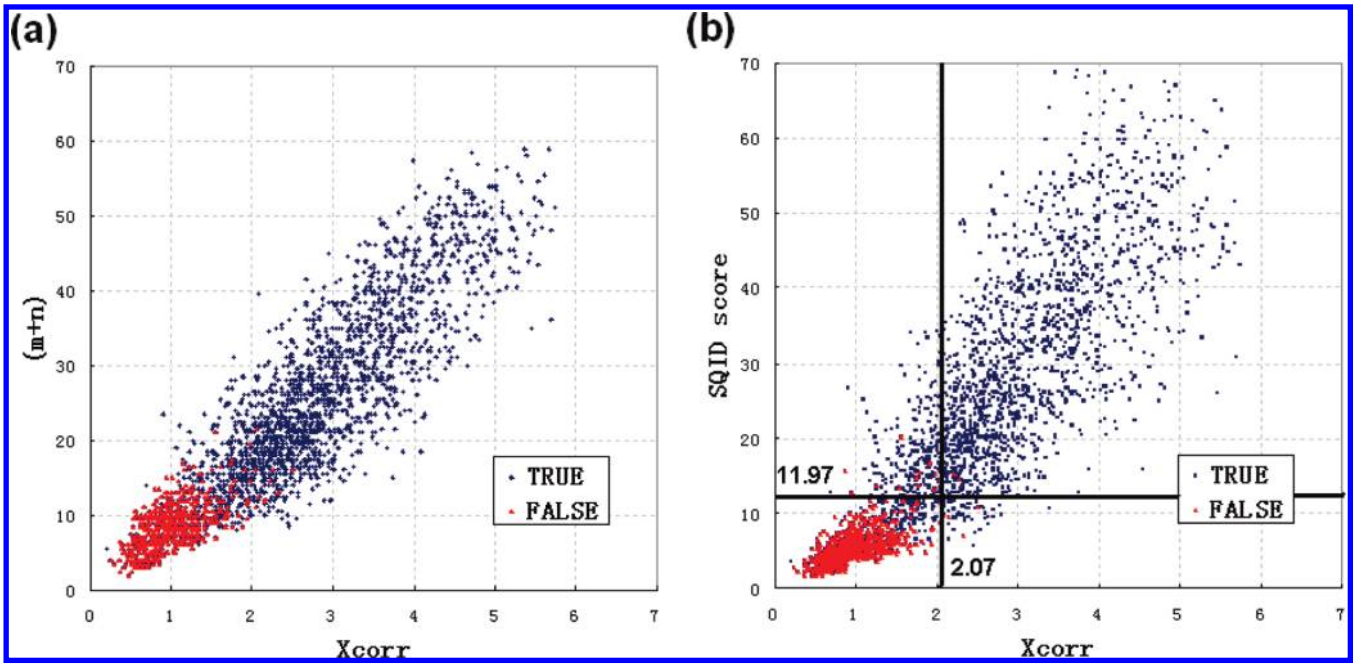


Figure 7. Plot of Xcorr versus (a) $m + n$ (numbers of matched peaks and numbers of consecutive pairs) and (b) SQID score for 2571 peptide-spectrum matches extracted from the 18 protein mixture data set. Every data point is scored by Sequest and SQID using the same experimental spectrum and the same peptide sequence. The blue spots are true identifications and red spots are false identifications.

are matching the same peptide sequence to the same spectra. We extracted the doubly charged spectra that reached the same peptide identifications by SQID and Sequest in the 18 protein mixture data set. Of the 18496 spectra of doubly charged peptides, only 2571 (13.9%) identifications were overlapping by SQID and Sequest, among which 1912 (74.4% of 2571) belonged to the 18 proteins or contaminations (True) and the remaining 659 belonged to reverse proteins (False). Figure 6a shows the plot of Xcorr with $(m + n)$, the SQID score without the intensity part, and Figure 6b plots Xcorr with the entire SQID score (with the intensity part incorporated). It can be seen that $(m + n)$ is almost proportional to Xcorr for both true and false identifications. However, the full SQID score increases much slower than Xcorr for false identifications in Figure 7b, and a better separation between true and false is achieved. The two bold lines are the corresponding Xcorr and SQID score thresholds (experimentally determined from Figure 4b) for 0.05 q-value cutoffs. At this confidence level, the peptides in the upper-right corner (Xcorr $\geq$ 2.07, SQID score $\geq$ 11.97) can be identified by both algorithms; the peptides in the upper-left corner (Xcorr < 2.07, SQID score $\geq$ 11.97) will be identified only by SQID and the peptides in the lower-right corner will be identified only by Sequest. Therefore, the plot suggests that combining different algorithms is potentially the most beneficial approach for maximizing the number of confident hits.

CONCLUSIONS

In general, SQID shows improved performance compared with popular algorithms as shown by results for three different data sets, with a large number of unique identifications. Combining SQID with other algorithms will thus be potentially beneficial, such as increasing the number of peptide hits and the confidence of identifications. SQID also has the potential to be applied to electron transfer dissociation (ETD) spectra as long as corresponding intensity tables are elucidated. By analyzing over 10 000 high resolution ETD spectra from the Coon group, the University of Wisconsin-Madison, we found that the peak intensities in ETD spectra are also highly dependent on the amino acid composition, for example, amino acid pairs containing basic residues tend to have enhanced cleavage, while pairs containing hydrophobic residues have weaker intensities. This study could help intensity prediction in ETD, and at the same time, provide evidence to clarify the controversial dissociation mechanisms. As a new algorithm, SQID still requires further optimization to improve the overall performance. Future efforts will include creating intensity histograms for other instrument types (e.g., Q-TOF, Orbitrap) and proteases (e.g., GluC, AspN, chymotrypsin) to enable SQID searches on these data, incorporating different intensity histograms corresponding to specific sequence motifs, combining SQID score and delta score to give a single discrimination score, and developing programs that directly modify scores from other search engines.

ASSOCIATED CONTENT

Supporting Information

Four supplementary tables, seven supplementary figures, and the intensity histogram for tryptic ion trap data that support the article are provided. This information is available free of charge via the Internet at http://pubs.acs.org. The SQID program and source code are under GNU GPL license and can be downloaded at http://quiz2.chem.arizona.edu/wysocki/bioinformatics.htm. The search results from SQID, Sequest and X!Tandem are available from the authors upon request.

AUTHOR INFORMATION

Corresponding Author

*Dr. Vicki H. Wysocki, Tel: 520-621-2628. Fax: 520-621-8407. E-mail: vwysocki@email.arizona.edu.

■ ACKNOWLEDGMENT

We appreciate NIST, PNNL, Dr. Andrew Keller, and Dr. Andrew Link for collecting and sharing the corresponding data sets for training and testing SQID. We also thank Dr. David Tabb for helpful discussions and suggestions. The study is funded by NIH Grant 2R01GM051387 to V.W.

■ REFERENCES

(1) Eng, J. K.; McCormack, A. L.; Yates, J. R. 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* (11), 976–989.

(2) Craig, R.; Beavis, R. C. Tandem: matching proteins with mass spectra. *Bioinformatics* **2004**, *20* (9), 1466–1467.

(3) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20* (18), 3551–3567.

(4) Sadygov, R. G.; Cociorva, D.; Yates, J. R. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat. Methods* **2004**, *1* (13), 195–202.

(5) Wysocki, V. H.; Tsaprailis, G.; Smith, L. L; Breci, L. A. Mobile and localized protons: a framework for understanding peptide dissociation. *J. Mass Spectrom.* **2000**, *35* (12), 1399–1406.

(6) Huang, Y.; Triscari, J. M.; Tseng, G. C.; Pasa-Tolic, L.; Lipton, M.. S.; Smith, R. D.; Wysocki, V. H. Statistical characterization of the charge state and residue dependence of low-energy CID peptide dissociation patterns. *Anal. Chem.* **2005**, *77* (18), 5800–5813.

(7) Huang, Y.; Tseng, G. C.; Yuan, S.; Pasa-Tolic, L.; Lipton, M. S.; Smith, R. D.; Wysocki, V. H. A data-mining scheme for identifying peptide structural motifs responsible for different MS/MS fragmentation intensity patterns. *J. Proteome Res.* **2008**, *7* (1), 70–79.

(8) Paizs, B.; Suhai, S. Fragmentation pathways of protonated peptides. *Mass Spectrom. Rev.* **2005**, *24* (4), 508–548.

(9) Zhang, Z. Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. *Anal. Chem.* **2004**, *76* (14), 3908–3922.

(10) Gibbons, F. D.; Elias, J. E.; Gygi, S. P.; Roth, F. P. SILVER helps assign peptides to tandem mass spectra using intensity-based scoring. *J. Am. Soc. Mass Spectrom.* **2004**, *15* (6), 910–912.

(11) Havilio, M.; Haddad, Y.; Smilansky, Z. Intensity-based statistical scorer for tandem mass spectrometry. *Anal. Chem.* **2003**, *75* (3), 435–444.

(12) Zhang, Z. Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.* **2004**, *76* (14), 3908–3922.

(13) Elias, J. E.; Gibbons, F. D.; King, O. D.; Roth, F. P.; Gygi, S. P. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat. Biotechnol.* **2004**, *22* (2), 214–219.

(14) Narasimhan, C.; Tabb, D. L; Verberkmoes, N. C.; Thompson, M. R.; Hettich, R. L.; Uberbacher, E. C. MASPIC: intensity-based tandem mass spectrometry scoring scheme that improves peptide identification at high confidence. *Anal. Chem.* **2005**, *77* (23), 7581–7593.

(15) Sun, S.; Meyer-Arendt, K.; Eichelberger, B. Improved validation of peptide MS/MS assignments using spectral intensity prediction. *Mol. Cell. Proteomics* **2007**, *6* (1), 1–17.

(16) Zhou, C.; Bowler, L. D.; Feng, J. A machine learning approach to explore the spectra intensity pattern of peptides using tandem mass spectrometry data. *BMC Bioinform.* **2008**, *9*, 325.

(17) Stein, S. E.; Pudnick, P. A. NIST Peptide Tandem Mass Spectral Libraries. Yeast Peptide Mass Spectral Reference Data, Yeast, ion trap, Official Build Date: May. 14, 2009. Drosophila Peptide Mass Spectral Reference Data, Drosophila, ion trap, Official Build Date: July. 14, 2008. National Institute of Standards and Technology, Gaithersburg, MD, 20899. Downloaded from http://peptide.nist.gov on April 1, 2010.

(18) Searle, B. C. Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics* **2010**, *10* (6), 1265–1269.

(19) Lipton, M. S.; Pasa-Tolic, L.; Anderson, G. A.; Anderson, D. J.; Auberry, D. L.; Battista, J. R.; Daly, M. J.; Fredrickson, J.; Hixson, K. K.; Kostandarithes, H.; Masselon, C.; Markillie, L. M.; Moore, R. J.; Romine, M. F.; Shen, Y.; Stritmatter, E.; Tolic, N.; Udseth, H. R.; Venkateswaran, A.; Wong, K.; Zhao, R.; Smith, R. D. Globalanalysis of the *Deinococcus* radiodurans proteome by using accurate mass tags. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (17), 11049–11054.

(20) Kolker, E.; Picone, A. F.; Galperin, M. Y.; Romine, M. F.; Higdon, R.; Makarova, K. S.; Kolker, N.; Anderson, G. A.; Qiu, X.; Auberry, K. J.; Babnigg, G.; Beliaev, A. S.; Edlefsen, P.; Elias, D. A.; Gorby, Y. A.; Holzman, T.; Klappenbach, J. A.; Konstantinidis, K. T.; Land, M. L.; Lipton, M. S.; McCue, L.; Monroe, M.; Pasa-Tolic, L.; Pinchuk, G.; Purvine, S.; Serres, M. H.; Tsapin, S.; Zakrajsek, B. A.; Zhu, W.; Zhou, J.; Larimer, F. W.; Lawrence, C. E.; Riley, M.; Collart, F. R.; Yates, J. R.; Smith, R. D.; Giometti, C. S.; Nealson, K. H.; Fredrickson, J. K.; Tiedje, J. M. Global profiling of *Shewanella oneidensis* MR-1: expression of hypothetical genes and improved functional annotations. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (6), 2099–2104.

(21) Keller, A.; Purvine, S.; Nesvizhskii, A. I.; Stolyar, S.; Goodlett, D. R.; Kolker, E. Experimental protein mixture for validating tandem mass spectral analysis. *OMICS* **2002**, *6* (2), 207–212.

(22) Zhang, B.; Chambers, M. C.; Tabb, D. L. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J. Proteome Res.* **2007**, *6* (9), 3549–3557.

(23) Tabb, D. L.; Fernando, C. G.; Chambers, M. C. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. Proteome Res.* **2007**, *6* (2), 654–661.

(24) Storey, J. D.; Tibshirani, R. Statistical significance for genome-wide studies. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (16), 9440–9445.