Published in final edited form as:

J Proteome Res. 2007 February ; 6(2): 654–661. doi:10.1021/pr0604054.

MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis

David L. Tabb^{2,1}, Christopher G. Fernando³, and Matthew C. Chambers²

² Mass Spectrometry Research Center/Departments of Biomedical Informatics and Biochemistry, Vanderbilt University Medical Center, 37232-8575

³ Engineering/Industrial Technology, West Virginia University Institute of Technology, 25136

Abstract

Shotgun proteomics experiments are dependent upon database search engines to identify peptides from tandem mass spectra. Many of these algorithms score potential identifications by evaluating the number of fragment ions matched between each peptide sequence and an observed spectrum. These systems, however, generally do not distinguish between matching an intense peak and matching a minor peak. We have developed a statistical model to score peptide matches that is based upon the multivariate hypergeometric distribution. This scorer, part of the “MyriMatch” database search engine, places greater emphasis on matching intense peaks. The probability that the best match for each spectrum has occurred by random chance can be employed to separate correct matches from random ones. We evaluated this software on data sets from three different laboratories employing three different ion trap instruments. Employing a novel system for testing discrimination, we demonstrate that stratifying peaks into multiple intensity classes improves the discrimination of scoring. We compare MyriMatch results to those of Sequest and X!Tandem, revealing that it is capable of higher discrimination than either of these algorithms. When minimal peak filtering is employed, performance plummets for a scoring model that does not stratify matched peaks by intensity. On the other hand, we find that MyriMatch discrimination improves as more peaks are retained in each spectrum. MyriMatch also scales well to tandem mass spectra from high-resolution mass analyzers. These findings may indicate limitations for existing database search scorers that count matched peaks without differentiating them by intensity. This software and source code is available under Mozilla Public License at this URL: <http://www.mc.vanderbilt.edu/msrc/bioinformatics/>.

Keywords

Proteomics; Identification; Statistical Distribution; Reversed Database; Peak Filtering

Introduction

Database search algorithms are the key bioinformatic tools to link proteomic tandem mass spectra to peptide sequences from protein databases. As shotgun proteomic technologies have developed, implementations such as Sequest¹ and Mascot² have been applied to ever larger collections of tandem mass spectra, emphasizing rapidity of identification. As proteomic technologies have been introduced to clinical applications, the accuracy of these algorithms has been tested by large protein sequence databases and aggressive post-translational modification (PTM) searches. The importance of scoring matches between peptide sequences

¹ Corresponding author: (phone) 615-936-0380; (fax) 615-343-8372; (email) david.l.tabb@vanderbilt.edu.

and MS/MS spectra can be observed in the diversity of algorithms created for this purpose. Comparisons have been conducted by cross correlation¹, hypergeometric distributions³, Poisson distributions⁴, Mowse scores², Bayesian statistics⁵, and dot products⁶, among others.

Evaluating the match between a candidate sequence and an observed tandem mass spectrum can be accomplished by examining the m/z locations that should be occupied by fragment ions from the sequence. In many database search algorithms, the count of these locations that contain peaks is a key factor in scoring the match through a statistical model. The most commonly used software of this type is Mascot², but others such as OMSSA⁴, RMET-Luck³, and PepProbe⁷ also employ this system to separate peaks into signal and noise. X!Tandem⁶ records the number of matched N-terminal and C-terminal fragments separately, combining this information with a dot product comparison of observed and expected intensities. In many cases, peak filtering takes place prior to matching in order to remove noise peaks from the spectrum. This filtering step, however, can introduce a problem. If an algorithm filters too aggressively, a peak representing an observed fragment ion may be lost; if it filters too few peaks, a noise peak may be erroneously matched to a fragment m/z projected from the peptide sequence.

Preprocessing of spectra is handled quite differently by the database search tools currently in use. Sequest keeps only the 200 most intense peaks for each spectrum and separates them into ten zones by their m/z ratios. Within each zone, Sequest normalizes the peak intensities to the highest peak of the zone, making the observed spectrum resemble the theoretical spectra to which it will be compared. Mascot and OMSSA, on the other hand, prioritize peaks by their intensities for matching to the theoretical peak lists; they progressively use larger collections of peaks for matching until scoring maximizes. By default, X!Tandem will employ the most intense 50 peaks from each ion trap tandem mass spectrum for matching. In short, essentially every database search engine focuses attention on the most intense peaks, and in most cases, peaks that do not make the cut are treated as if they were not observed.

The MASPIC scoring system⁸ separates peaks into multiple classes on the basis of intensity. The highest intensity class is restricted to very few peaks, whereas the remaining classes contain greater numbers of lower intensity peaks. Matching a peak from a class that contains few peaks contributes more to a high score than matching a peak from a more populous class. For each matched peak, MASPIC evaluates the intensity class and also the m/z fidelity (proximity of the expected and observed m/z values) to build a match probability through an application of the multinomial distribution. MASPIC conflates information about these two properties, making it unclear which is contributing most to discrimination. Further, its use of the multinomial distribution does not reflect that once a peak is matched from an intensity class, that peak is unavailable for matches to other predicted peaks for a particular candidate sequence, thus reducing the probability that further matches will occur from this class.

In this work, we examine MyriMatch, a database search algorithm that employs multiple peak intensity classes to estimate the probability of a match occurring by random chance. We have introduced two tunable elements in scoring. First, the preprocessing of spectra can be scaled to retain more or less of each spectrum's original intensity. Second, peaks are stratified into a configurable number of intensity classes. We employed the multivariate hypergeometric distribution to compute the probability of random match for each pairing of candidate sequence and spectrum. We evaluated the algorithm on data sets from three different labs and three different instruments to improve the generality of our conclusions. We compared its effectiveness to Sequest and X!Tandem, two widely used database search tools. Our results indicate that stratifying peak matches by intensity produces substantial gains in discrimination. By greatly reducing the need for peak filtering, MyriMatch can achieve greater sensitivity of identification than other algorithms. We also demonstrate its effectiveness for high-resolution

tandem mass spectra. MyriMatch and its accompanying tools are freely available with source code.

Experimental

Data sources

Because tandem mass spectra reflect the instrument configuration that produced them, we sought data sets from a variety of sources to generalize the assessment of the MyriMatch algorithm. Shotgun proteomics experiments⁹ were performed in three different laboratories on three different instruments:

- In the Banting and Best Department of Medical Research at the University of Toronto, the cytosolic proteins of mouse placentas were reduced and alkylated, denatured, and digested with trypsin¹⁰. The resulting peptides were separated in a 100 μ m ID SCX (strong cation exchange)/RP (reversed phase) biphasic column for a 12-cycle MudPit on a Thermo Finnigan LCQ Deca XP 3D ion trap tandem mass spectrometer (Thermo Finnigan, San Jose, CA). Three microscans were summed to produce each full scan, and two microscans were summed to produce each tandem mass spectrum. We selected the seventh cycle of the MudPit to evaluate these algorithms due to its high peptide content. Spectra were identified with the International Protein Index (IPI) database for mice, version 3.22 (51,477 proteins).
- In the Clinical Pharmacology Division and in the Proteomics Laboratory of the Mass Spectrometry Research Center at Vanderbilt University, human atrial cells were lysed. A fusion protein of GST and the C-terminus of a cardiac potassium channel (KCNA5) was employed to pull down interacting protein partners. The proteins were eluted, reduced, and alkylated prior to trypsin digestion. The resulting peptides were subjected to RP LC (liquid chromatography) in a 100 μ m column into a Thermo Finnigan LTQ linear ion trap tandem mass spectrometer. Each tandem mass spectrum consisted of a single microscan. Spectra were identified with the human IPI database, version 3.22 (57,846 proteins).
- Also in the Proteomics Laboratory, a defined mixture of 49 human proteins from Sigma Aldrich (St. Louis, MO) was reduced and alkylated prior to trypsin digestion. The resulting peptides were subjected to RPLC in a 100 μ m column into a Thermo Finnigan LTQ linear ion trap tandem mass spectrometer. Each tandem mass spectrum consisted of a single microscan. Spectra were identified in a database containing the 49 known protein sequences and 76,229 decoy sequences from the zebrafish and *Arabidopsis* IPI databases.
- In the Organic and Biological Mass Spectrometry Group at Oak Ridge National Laboratory, *Rhodospseudomonas palustris* lysates were fractionated by ultracentrifugation into membrane and crude fractions¹¹. The crude fraction was denatured and reduced prior to trypsin digestion. A 100 μ m biphasic (SCX/RP) column separated the resulting peptides for three different experiments. First a Thermo Finnigan LTQ linear ion trap tandem mass spectrometer was used. Second, a Thermo Finnigan Orbitrap collected full scans in the Orbitrap and tandem mass spectra in the LTQ mass analyzer. Third, a Thermo Finnigan Orbitrap collected both full and tandem mass spectra in the Orbitrap mass analyzer. In all cases, two microscans were summed for each full scan and for each tandem mass spectrum. We selected the second cycles of these MudPits for this experiment. Spectra were identified with an *R. palustris* sequence database that was provided by ORNL (4848 proteins).

All database searches (except for the defined 49 protein mixture) were against databases that contained each protein sequence in both normal and reversed orientations, doubling the above numbers of proteins for each database. Candidate peptides were required to be tryptic on both ends. Oxidation of methionine was considered as a possible PTM.

Database search software

The “MyriMatch” database search software¹ was created in the C++ language to identify peptides with high discrimination and high speed. The software is multi-threaded to take advantage of multiple processors and cores, and it can distribute processing over clusters by use of MPI (Message Passing Interface). The software reads tandem mass spectra from files in the mzData 1.05 format¹². It separates spectra into those from singly-charged precursors and those from multiply-charged precursors. The latter are identified twice, once assuming a +2 charge state and once assuming a +3 charge state.

Tunable preprocessing and scoring—During preprocessing, the software computes the total ion current (TIC) for each spectrum by summing the fragment ion intensities for each MS/MS. Users specify the proportion of the TIC to be retained during preprocessing. The software sorts the peaks in descending order of intensity and retains the minimum number required to meet this target TIC (see Figure 1). By removing a percentage of TIC rather than a percentage of peaks, MyriMatch produces a final peak list that reflects the distribution of intensity in the original peaks; spectra in which the signal is spread over many fragments retain a higher percentage of peaks than spectra dominated by a small number of intense peaks.

Users also specify the number of intensity classes that should be employed during scoring. These classes differ in the number of peaks they hold so that adjacent class sizes are in a ratio of 2:1, with the most intense class holding the fewest peaks². If three intensity classes are selected and 70 peaks are retained from the MS/MS, the most intense “A” set will contain 10 peaks, the intermediate “B” set will contain 20 peaks, and the least intense “C” set will contain 40 peaks. If a spectrum contains too few peaks to fill the intensity classes, the spectrum is bypassed. When using three classes, for example, seven peaks are required for each spectrum (1+2+4).

Fragmentation model—Candidate peptides are generated from a protein sequence database in simultaneous linear scan of the database sequences^{13,14}. A candidate can be compared to a spectrum if it falls within the m/z tolerance of the intact peptide ion (1.25 m/z from average mass for ion trap full scans, 0.05 m/z from monoisotopic mass for Orbitrap full scans). To compare a peptide to a spectrum, the software generates the m/z positions at which it expects to observe fragment ions for the sequence. For singly- and doubly-charged precursors, singly-charged b and y ions are modeled. In the case of triply-charged precursors, MyriMatch employs a novel system for modeling fragments. Each residue in the peptide is given a weight; Arg, His, and Lys are given a weight of five, Gln and Asn are given a weight of three, and other residues are assigned a weight of one. For each peptide bond, the summed weight on both sides is computed, and the terminus with the higher weight is generated as a doubly-charged fragment ion while the other terminus is generated as a singly-charged fragment ion. By this mechanism, the system models peptides from all three charge states with two fragment ions from each peptide bond, making the resulting probabilities more comparable.

¹Because this software was designed for multi-core CPU and multi-computer configurations, its name is derived from the word *myriad*, meaning “countless or innumerable.”

²This ratio was chosen as the smallest integer that would differentiate the class populations. A system to dynamically populate classes to reflect the distribution of spectral intensity discriminated sequences less well (data not shown).

Match scoring—For each predicted m/z value, the corresponding location in the spectrum is tested to determine whether or not any peak is present, and if so, what class it is. To match, an observed peak must fall within a specified tolerance of the target location (0.5 m/z for ion traps, 0.03 m/z for the Orbitrap). If multiple peaks fall in this range, the one nearest the expected location is taken as the match. The match between the candidate sequence and the spectrum can be represented by the number of missing peaks along with the number of matches in each intensity class. For each spectrum, the software has pre-counted the total number of locations at which no peak is present as well as the total number of peaks in each intensity class. For instruments of greater mass accuracy, the number of locations that can be occupied for a given range of m/z values is higher than for ion traps. This total number of locations can be envisioned as a collection of balls. White balls represent those locations that are unoccupied by peaks. Red balls represent class A peaks, twice as many green balls represent class B peaks, and twice again as many blue balls represent class C peaks. Each match can then be represented by selecting a number of balls from this collection.

To compute the probability of this match occurring by random chance, the software employs

$$p = \frac{\prod \binom{t_i}{m_i}}{\binom{T}{M}}$$

the multivariate hypergeometric (MVH) distribution: $\binom{t_i}{m_i}$, where p is the probability of the match occurring by random chance, t_i is the number of peaks from a particular intensity class in the spectrum, m_i is the number of peaks from a particular intensity class matched to the peptide-derived peak-list, T is the total number of locations in the spectrum, and M is the total number of peaks predicted from the peptide sequence. Note that the numbers of missed matches and of “missing” peaks in the spectrum must be included among the t_i and m_i for this probability to be correct. For example, one spectrum contains 82 class A peaks, 164 class B peaks, and 328 class C peaks (red, green and blue balls). It extends from 250 to 1775 m/z ; the number of locations that can be occupied in the spectrum (total number of balls) is 1525, of which 574 are occupied by classified peaks, leaving 951 voids (white balls). Of 32 fragment ions that were predicted for a particular peptide sequence, one fell outside the scan range and was disregarded. For the remaining 31 m/z values, 21 matched to class A peaks, 6 matched to class B peaks, and 1 matched to a class C peak, leaving 3 unmatched. The probability for this

$$\text{match is } \frac{\binom{82}{21} \binom{164}{6} \binom{328}{1} \binom{951}{3}}{\binom{1525}{31}} \text{ or } 4.78 \times 10^{-25}.$$

Computing combinations in this form

is problematic since each combination implies performing three factorials: $\binom{a}{b} = \frac{a!}{(a-b)!b!}$. Instead, MyriMatch generates a lookup table for the first few thousand values of $\ln(c!)$ and

$$\ln \binom{a}{b} = \ln(a!) - \ln((a-b)!) - \ln(b!)$$

computes natural logarithms of combinations:

Transformed to the log domain, the MVH distribution can be calculated with only additions

$$\ln(p) = \left[\sum \ln \binom{t_i}{m_i} \right] - \ln \binom{T}{M}$$

and subtractions: The negative of this log probability is reported; the score for the above example is 56, indicating that the probability of this match occurring at random is e^{-56} . When matching is complete, the five best scoring sequences are recorded for each spectrum to a unified SQT file¹⁵ to preserve compatibility with DTASelect¹⁶.

MyriMatch employs the MVH distribution rather than the multinomial distribution⁸ because matching to a peak of the highest intensity class reduces the probability that another peak will

be randomly matched from this class by other predicted peaks; this is equivalent to the “no replacement” rule that differentiates the MVH distribution from the multinomial distribution. To return to the ball analogy, this means that once a red ball has been drawn from the collection, the probability of drawing a red ball has been reduced because that ball has been selected and removed from the collection. Because each observed peak can only be matched to a single expected m/z value from a candidate peptide sequence, the same replacement rule applies. The probabilities derived in this way from the MVH distribution are very similar to those produced by RMET-Luck³ with two key exceptions. First, they are computed by the multivariate hypergeometric rather than the univariate hypergeometric distribution (to differentiate peak matches by intensity). Second, the match probability is used without normalization; RMET-Luck divides the match probability by the mode probability, but multiple modes may exist in the MVH distribution.

Peptide Identification Format Conversion

As a basis for comparison, we identified spectra with two other database identifiers. The primary identifier in use at Vanderbilt is the Sequest algorithm (Thermo Electron Corporation, San Jose, CA) version 27, revision 12¹. We also employed X!Tandem 2, version 2006.09.15.1⁶, an open-source identifier that has rapidly grown in popularity. X!Tandem can potentially refine its identifications by searching for semi-tryptic peptides and PTMs in a second pass; we configured the software to skip this step because this research was intended to test scorers, not the frameworks in which they are situated. We found, however, that X!Tandem continued to apply certain N-terminal modifications to its search. As a result, X!Tandem has a small advantage over Sequest and MyriMatch in this comparison. We elected not to use Mascot from Matrix Science (London, UK) because the version licensed to Vanderbilt has been superseded. Because Sequest produces its identifications in thousands of OUT files and X!Tandem produces identifications in an XML file containing many other types of information, this comparison required the transcoding of identifications from these algorithms into the SQT format¹⁵ employed by MyriMatch.

The “SQTer” tool was developed in C++ to create SQT files from Sequest OUT files and X!Tandem XML results. In the case of Sequest results, the one-to-one correspondence between OUT and SQT formats makes this purely a transcoding operation. X!Tandem results, on the other hand, have both more and less information than Sequest results. The results, for example, include the alignment of each identified peptide to every protein in which it is found. On the other hand, X!Tandem does not report matching sequences that rank below the best match for each spectrum, and it reports only those matches that have some chance of being correct. SQTer records the hyperscore for each peptide match in the SQT “XCorr” field, and it records the reciprocal of the expectation value in the “DeltCN” field (to retain the property that a high score indicates greater reliability). This placement reflects that fact that the X!Tandem hyperscore, like XCorr, reflects the quality of the match without regard for other potential matches, while DeltCN and expectation values are relative score evaluations that reflect the separation of the best matches from other identifications. By converting Sequest and X!Tandem results to the format employed by MyriMatch, these algorithms can be compared on an even basis.

Evaluation of peptide identification

The standard way to evaluate database search scorers has been the construction of ROC (receiver operator characteristic) plots for samples of defined content. Identifications are labeled true if they are made to proteins known to be in the mixture and false if they are made to proteins in a larger decoy set. Here, we present a new way to make this evaluation based on identification confidences based on reversed database searching^{17–19}. Identification for complex samples is performed against protein databases containing both normal and reversed

versions of each sequence. The resulting identifications are then filtered so that a specified level of confidence is achieved (equivalent to requiring a fixed ratio of forward to reversed peptide identifications). This filter is based upon on *peptide* identifications, not *protein* identifications; the confidence of protein identifications will vary depending on the rules for filtering at this level, such as removal of single-peptide proteins. By comparing the number of passing peptide identifications from different algorithms or different configurations, one compares the extent to which the algorithm can elevate the scores of legitimate identifications while reducing the scores of reverse identifications. As a result, samples of greater complexity and relevance can be used for evaluation.

Software entitled “DBValidate” was created in the C++ language to determine identification score thresholds that correspond to a user-specified confidence. The software is intended for use on SQT files from database searches employing databases containing both normal and reversed sequences. Identifications from each search are subdivided by charge state. Each identification is classified “real” if the peptide comes from a protein of forward (normal) orientation, “decoy” if the peptide comes from a reversed protein sequence, or “both” if the peptide can be found in proteins of both orientations. DBValidate examines the peptide hits at a series of thresholds to determine the number of real and decoy peptides passing the threshold. It estimates the number of erroneous identifications to be double the number of passing decoy peptides, and it estimates the number of correct identifications to be the number of passing real peptides minus the number of decoy peptides. The confidence of the passing peptides is then estimated by dividing the number of estimated correct identifications by the total number of passing real and decoy peptides. When the ratio of real to decoy sequences is not 1:1 (either because a subset of sequences has been reversed or because a larger decoy database has been employed), DBValidate uses the ratio of real to decoy sequences as a multiplier rather than simply doubling the number of passing “decoy” peptides.

DBValidate can sort identifications by either the absolute or relative scores recorded in the SQT files. The absolute scores for each of the three algorithms include the MyriMatch random match probability, the Sequest cross correlation score (XCorr), and the X!Tandem hyperscore. The DeltCN scores for MyriMatch and Sequest were computed in the same way; the second-best score was subtracted from the top score, and this difference was normalized by the best score. X!Tandem computes expectation values in a more complex way, approximating the fit of the top score into the distribution of lower scores. Our standard threshold was 95% confidence. At this level, the mix of peptide identifications passing the thresholds should be 19 correct peptide identifications for each false one. This system for thresholding these identifications can counterbalance the effects of larger sequence databases and the use of PTMs in searching. By learning thresholds from the data, DBValidate can level the playing field for disparate database search strategies.

Results and Discussion

Spectral differences and preprocessing

The tandem mass spectra from the six evaluated samples are summarized in Table 1. To survey these data is to revisit a history of progress in shotgun proteomics instrumentation. The LCQ 3D ion trap acquired approximately 25 spectra a minute. Its successor, the LTQ linear ion trap, improved on this scan rate by more than a factor of five. Improvements are not limited to numbers of spectra, either; the Orbitrap enables the collection of high resolution tandem mass spectra at a rate higher than that of the LCQ. The diversity of these samples reveals that instrument configuration and sample complexity can change the resulting spectra considerably. Although the human and standard protein mixtures were run under the same configuration, the spectra from the human sample contained far larger numbers of peaks. When an LTQ configured for two microscans per ms/ms analyzed the *R. palustris* sample, the number of

peaks per spectrum increased dramatically. In fact, the interquartile ranges (the middle 50%) of the spectral peak counts in the *R. palustris* and human samples do not overlap, despite their having been collected on the same type of instrument. When examining the spectra from a particular run, it is generally the case that spectra at the 75th percentile contain twice as many peaks as the spectra at the 25th percentile. It should be apparent that treating all tandem mass spectra as interchangeable may lead to a loss of accuracy in identifying them.

Figure 2 shows the effects of filtering peaks to retain a percentage of the TIC for each spectrum. Each curve corresponds to a spectrum from a different sample. The chosen spectrum for each sample was the one at the 75th percentile of original peak counts (the third quartile was chosen because identified spectra tend to have higher peak counts). Reducing retained TIC by a small percentage reduces the number of retained peaks by a higher percentage because small peaks are far more numerous than intense peaks. When removing only two percent of spectral TIC, MyriMatch typically removes between ten and twenty-five percent of the peaks. By removing a fraction of TIC rather than a fraction of peaks, MyriMatch filters a higher percentage of peaks from dense spectra than from sparse spectra, making them more uniform in composition. The higher prevalence of small peaks underlies the scoring model of MyriMatch in which the low-intensity classes contain geometrically more peaks than the higher-intensity classes.

Peak Filtering and Intensity Classification

Spectra from each sample were subjected to MyriMatch search 64 times. Four intensity class configurations were tested, from a one-class match/no match system up to a four-class model which segregated observed peaks into four levels of intensity. Each increase in class number placed an increasing amount of importance on the intensity information for these peaks. Each intensity class configuration was tested sixteen times, where the percentage of retained TIC ranged from 70% to 100% in steps of 2%. The number of identifications for each of these 64 runs is represented by a point on the lines shown in Figure 3a–c. The numbers of identifications are reported in Supporting Information along with the graphs for all six of the samples.

An examination of the results from the LCQ sample in Figure 3a reveals that filtering at different levels of TIC has non-uniform effects. The number of identifications varies with even minor differences in peak filtering. This reflects that the spectra for this RPLC run are heterogeneous, requiring different amounts of filtering for optimal identification. The general rise of the lines from left to right, however, reflect that the proportion of identifiable spectra is higher when more of the original peaks are retained for matching. The exception to this improvement is the red line, which represents a scoring model where expected peaks are either matched or not matched. This simple scoring model saturates when all peaks are retained, and its discriminating power diminishes as a result. The maximum number of identifications for this simple model falls behind the overall maximum by 11%. If intensity is used to stratify the observed peaks into separate classes, however, the LCQ spectra can be identified without peak filtering, maximizing sensitivity and specificity of identification.

The LTQ results for the human heart immunoprecipitation sample (Figure 3b) are representative of the other LTQ samples. If all peaks are retained for matching, the performance of the match/no match model plummets. At best, its performance lags the overall maximum by 7.5%. The two-class model, which separates observed peaks into intense and minor peaks, becomes saturated at high percentages of peak retention as well. The models that employ three- and four-way classification, however, are resistant to this discriminative degradation.

The trends observed in tandem mass spectra collected in an Orbitrap (Figure 3c) are considerably different from those observed for the LCQ and LTQ. In this evaluation, the number of intensity classes is largely irrelevant. The only overall trend is that retaining more peaks leads to better discrimination. What sets these spectra apart is that fragment ions are

considered matches if they fall within 0.03 Da of their expected locations. This means that each spectrum has a far larger number of potential locations for peaks, and yet Table 1 shows that these spectra were among the most sparsely populated with peaks. The candidates can be separated based on the count of matched fragment ions without the need to rely on intensity as an additional discriminator.

Overall performance comparison

These results informed our choice of settings for routine MyriMatch use. By retaining 98% of spectral intensity, MyriMatch is able to remove obvious noise peaks from spectra without reducing discrimination in most cases. We opted to stratify the remaining fragment ions into three classes; this model is effective in very dense and sparse spectra alike. Splitting to four classes appears to pose no advantage over division into three classes. The configuration of 98% TIC at three classes was not uniformly superior in performance, but it was consistently within 5% of the best observed performance for each sample.

Figure 3d compares the performance of MyriMatch to that of Sequest and X!Tandem. Probabilities of random matching were used for confidence filtering in MyriMatch results. Sequest identifications were sorted by XCorr for filtering, while X!Tandem peptide identifications were sorted by their expectation values for thresholding. Filtering was based entirely on the ability of the identifiers to discriminate matches to real protein sequences from those to decoy proteins.

MyriMatch yielded the largest number of confident identifications in five of the six samples. Sequest achieved the best discrimination for the LCQ sample. In general, the three yielded similar performance. The biggest performance difference came from the LCQ sample, where the worst performance trailed the best by 27%. The smallest difference (6%) was found in the Orbitrap-LTQ analysis of the *R. palustris* sample. This small difference may reflect that the precursor masses were known very accurately from the Orbitrap MS scans, so approximately 300 candidate sequences were compared to each spectrum (note also that this microbial sequence database is quite small). With fewer candidate sequences proposed for each spectrum, the difficulty of discerning the correct match was reduced.

Each algorithm has strengths and weaknesses. In contrast to its first-ranked performance on LCQ data, Sequest lagged 16% behind MyriMatch in the Orbitrap tandem mass spectra because the version licensed to Vanderbilt could not take the high mass accuracy of the fragments into account. X!Tandem produced very similar performance to MyriMatch for most of the samples using LTQ spectra, but its performance lagged by almost a quarter in the LTQ *R. palustris* sample (where the spectra were particularly dense with peaks). This drop in performance was not seen in the results for X!Tandem in the Orbitrap/LTQ results for this sample. This may reflect the much reduced pool of candidate sequences resulting from the high resolution MS data. Researchers should ensure that their data collection approaches are playing to the strengths of their identification tools.

Peptide overlap among identifiers

We focused on the sample with defined content to compare the peptides identified by the three algorithms. The sample contained 49 proteins from a Sigma human protein standard. DBValidate filtered the peptide identifications for each of the three algorithms, and the passing peptide sequences were enumerated in each case. For Sequest, this filtering produced XCorr thresholds of 1.92 (singly-charged), 2.76 (doubly-charged), and 3.00 (triply-charged). The thresholds for X!Tandem were expectation values of 0.035 (singly-charged), 0.04 (doubly-charged), and 0.025 (triply-charged). MyriMatch negative log probabilities were filtered at these levels: 29.2 (singly-charged), 28.0 (doubly-charged), and 32.2 (triply-charged).

Redundant appearances of each peptide may have resulted from duplicate spectra or due to spectra from multiple precursor charge states. These redundancies were removed so that each sequence could appear just once on the list for each algorithm. These lists were compared to produce a Venn diagram (see Figure 4).

The overlap between pairs of algorithms reveal that MyriMatch and X!Tandem are the most similar in the peptides they identify. Of the peptide identifications from either MyriMatch or X!Tandem, 68% were identified by both algorithms. This percentage was 61% for Sequest and X!Tandem and 62% for Sequest and MyriMatch. The similarity in identified peptides between X!Tandem and MyriMatch may reflect the relationship between their scorers; both are seeking peaks at particular m/z values and recording the result. Sequest, on the other hand, is comparing the frequencies of the observed spectrum to those of the modeled spectra, an approach that is orthogonal to the peak lookup strategy. 51% of the peptides were identified by all three algorithms.

The lengths of peptide sequences were lower on average for peptides identified only by MyriMatch (11.2 residues) or by both MyriMatch and X!Tandem (10.6 residues). Peptides identified only by X!Tandem, on the other hand, averaged 14.5 residues, and peptides identified only by Sequest averaged 14.4 residues. Because shorter peptides yield spectra with fewer fragment ions, their identification by MyriMatch is evidence that it can make maximal use of minimally informative spectra.

A second Venn diagram was created using an identification confidence of 99% (not shown). The number of peptides identified overall dropped from 482 to 378 (-21.6%). Of the peptides identified by either MyriMatch or X!Tandem, 68% were again observed to be identified by both. The overlap in performance between X!Tandem and Sequest was also unchanged at 61%. The percentage of sharing between MyriMatch and Sequest, however, fell to 57% because of a reduction in both the percentage of peptides identified by all three algorithms (-2.7%) and the percentage of peptides identified only by Sequest and MyriMatch (-2.3%). In total, however, the overlap among the three algorithms changed very little as the peptide confidence filtering was increased in stringency.

Conclusions

MyriMatch is a powerful open-source database identifier. Its scoring system emphasizes matches to intense peaks over matches to minor peaks though an application of the multivariate hypergeometric distribution. Because this scoring scales to handle large numbers of peaks per spectrum, less peak filtering is required, enabling higher sensitivity of identification. In its current implementation, MyriMatch out-performs Sequest and X!Tandem in identifying peptides from reversed database searches.

Comparing peptide identifications from these three algorithms makes it apparent that applying different scoring systems can identify more diverse collections of peptides than applying a single scoring system. Moving forward, we intend to build alternative scoring systems into MyriMatch so that each candidate peptide is evaluated by a variety of orthogonal scoring systems. If successful, this approach will give the advantages of multiple search tools without requiring applications of different algorithms and conversion of their outputs. Here we have shown the application of the multivariate hypergeometric distribution to scoring matches on the basis of peak intensity. The same distribution could also be used to emphasize peaks that have lost water or ammonia neutrally²⁰. By scoring separate aspects of each match by separate scorers, MyriMatch may gain considerable discriminative power.

In this report, all six data sets were produced on Thermo mass spectrometers. They varied considerably, however, in the number of peaks contained by each spectrum. MyriMatch is

robust against these differences in peak density. Database identifiers that rely on match/no match scoring may not scale well to some data sets, such as those summing two microscans in the LTQ. The diversity of tandem mass spectrometry data has not been sufficiently recognized in the design of database search tools.

The software employed in this work is available with source code under the Mozilla Public License. The elements include ReAdW, which we modified to produce mzData files; MyriMatch, the C++ database identifier; SQTer, a tool to create SQT files from Sequest and X!Tandem outputs; and DBValidate, a system for inferring score thresholds that correspond to a particular false positive rate. The software is available at this URL: <http://www.mc.vanderbilt.edu/msrc/bioinformatics/>

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by NIH/NCI 1 R01 CA126218-01, NIH/NCI 1 U24 CA126479-01, NIH P30 ES000267, and by an American Cancer Society Institutional Research Grant (#IRG-58-009-48) through the Sartain-Lanier Family Foundation and VICC Discovery Grant Programs. Nathan VerBerkmoes, Manesh Shah, and the OBMS group of Oak Ridge National Laboratory graciously provided the data from *Rhodopseudomonas palustris* with support from GTL grant DOE DE-FG02-01ER63241. Kathy Murray of the Vanderbilt Cardiovascular Medicine Division helpfully provided the human heart immunoprecipitation samples with support from NIH/NHLBI HL071002. Amy Ham in the Vanderbilt Mass Spectrometry Research Center produced the mass spectrometry data from both the human heart and human protein standard mixture samples. Thomas Kislinger analyzed the mouse placentas in Andrew Emili's group at the University of Toronto and graciously provided the data for this work. DLT thanks Tema Fridman for her patient explanation of the RMET-Luck scoring model.

References

1. Eng JK, McCormack AL, Yates JR 3rd. J Am Soc Mass Spectrom 1994;5:976–989.
2. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Electrophoresis 1999;20:3551–67. [PubMed: 10612281]
3. Fridman T, Razumovskaya J, Verberkmoes N, Hurst G, Protopopescu V, Xu Y. J Bioinform Comput Biol 2005;3:455–76. [PubMed: 15852515]
4. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH. J Proteome Res 2004;3:958–64. [PubMed: 15473683]
5. Zhang N, Aebersold R, Schwikowski B. Proteomics 2002;2:1406–12. [PubMed: 12422357]
6. Craig R, Beavis RC. Bioinformatics 2004;20:1466–7. [PubMed: 14976030]
7. Sadygov RG, Yates JR 3rd. Anal Chem 2003;75:3792–8. [PubMed: 14572045]
8. Narasimhan C, Tabb DL, Verberkmoes NC, Thompson MR, Hettich RL, Uberbacher EC. Anal Chem 2005;77:7581–93. [PubMed: 16316165]
9. Washburn MP, Wolters D, Yates JR 3rd. Nat Biotechnol 2001;19:242–7. [PubMed: 11231557]
10. Kislinger T, Cox B, Kannan A, Chung C, Hu P, Ignatchenko A, Scott MS, Gramolini AO, Morris Q, Hallett MT, Rossant J, Hughes TR, Frey B, Emili A. Cell 2006;125:173–86. [PubMed: 16615898]
11. VerBerkmoes NC, Shah MB, Lankford PK, Pelletier DA, Strader MB, Tabb DL, McDonald WH, Barton JW, Hurst GB, Hauser L, Davison BH, Beatty JT, Harwood CS, Tabita FR, Hettich RL, Larimer FW. J Proteome Res 2006;5:287–98. [PubMed: 16457594]
12. Orchard S, Hermjakob H, Taylor CF, Potthast F, Jones P, Zhu W, Julian RK Jr, Apweiler R. Proteomics 2005;5:3552–5. [PubMed: 16167370]
13. Edwards N, Lippert R. Proceedings of the Second International Workshop on Algorithms in Bioinformatics 2002;2452:68–81.
14. Tabb DL, Narasimhan C, Strader MB, Hettich RL. Anal Chem 2005;77:2464–74. [PubMed: 15828782]

15. McDonald WH, Tabb DL, Sadygov RG, MacCoss MJ, Venable J, Graumann J, Johnson JR, Cociorva D, Yates JR 3rd. *Rapid Commun Mass Spectrom* 2004;18:2162–8. [PubMed: 15317041]
16. Tabb DL, McDonald WH, Yates JR 3rd. *J Proteome Res* 2002;1:21–6. [PubMed: 12643522]
17. Cargile BJ, Bundy JL, Stephenson JL Jr. *J Proteome Res* 2004;3:1082–5. [PubMed: 15473699]
18. Higdon R, Hogan JM, Van Belle G, Kolker E. *Omics* 2005;9:364–79. [PubMed: 16402894]
19. Moore RE, Young MK, Lee TD. *J Am Soc Mass Spectrom* 2002;13:378–86. [PubMed: 11951976]
20. Dancik V, Addona TA, Clauser KR, Vath JE, Pevzner PA. *J Comput Biol* 1999;6:327–42. [PubMed: 10582570]

Peak Intensity and Classification

Figure 1.

During preprocessing, the fragment ions for each spectrum are sorted by intensity in decreasing order. The cumulative intensity from the most intense peak to each other peak is computed (represented here by the curve). A fraction of the original TIC is retained (in this case, 95%), with the remaining peaks stripped out as noise. The remaining peaks are split into classes that double in population at each step; in this three class example, each is marked with a letter.

Figure 2.

Sample and instrument configuration can lead to very different numbers of peak per tandem mass spectrum. The LTQ produced spectra with the highest density and the lowest density. Tandem mass spectra from an Orbitrap will differ by the mass analyzer in which they were collected. In this figure, the number of peaks remaining after peak filtering is shown for the spectrum at the third quartile of peak counts. Peak filtering on the basis of TIC retention reveals the exponential distribution of peak intensities; the least intense peaks are most common, while intense peaks are least numerous. Filtering out only 8% of total intensity in the LTQ Human sample reduced peak counts by half.

LCQ Mouse placenta MudPit cycle

Orbi-Orbi R. palustris MudPit cycle

Algorithm performance across samples

Figure 3.

Figure 3a-3d. MyriMatch performance was evaluated on tandem spectra from LCQ, LTQ, and Orbitrap mass analyzers. Scoring models with one, two, three, and four intensity classes were tested. The higher the curve, the more peptides were confidently identified at this configuration of the software. The horizontal position reflects the extent of preprocessing, with the heaviest filtering on the left and the least filtering on the right. A scoring system that counts only matches and mismatches (represented by the red curves) performs worse with little peak filtering in LCQ and LTQ data. Only Orbitrap tandem mass spectra reveal no difference in performance when peaks are stratified by their intensities. In the final panel, the performance of MyriMatch (retaining 98% of and segregating peaks into three intensity classes) is compared to that of Sequest and X!Tandem for all six evaluated samples.

Figure 4.

Figure 4 (TOC Graphic). This Venn diagram shows the overlap in peptide identifications among the three tested algorithms for the human standard protein mixture. MyriMatch (394 peptides) and X!Tandem (369 peptides) outperformed Sequest (309 peptides) for this sample. The gained peptides for MyriMatch tended to be shorter sequences. Of all observed peptides, 51% were identified by all three algorithms.

Table 1

Sample	Analyzer		MS2	uScans	MS2Spectra	Spectra/min	Peak Counts	
	MS1						25%ile	75%ile
Mouse standard Human <i>R. palustris</i> <i>R. palustris</i> <i>R. palustris</i>	LCQ		LCQ	3/2	3106	24.8	182	334
	LTQ		LTQ	1/1	10862	141.1	58	200
	LTQ		LTQ	1/1	10981	142.6	154	578
	LTQ		LTQ	2/2	8290	69.1	605	1167
	Orbi		LTQ	2/2	7715	64.3	571	1127
	Orbi		Orbi	2/2	4465	37.2	109	210

The six data sets used in this study were collected in three different laboratories using LCQ, LTQ, and Orbitrap instruments. Using multiple LTQ microscans led to a slower rate of ms/ms acquisition and produced spectra that were dense with peaks. The Orbitrap was faster collecting ms/ms spectra than the LCQ, and spectra were relatively sparse. Spectra from different configurations of the LTQ ranged considerably in the numbers of peaks they contained. The spectra from each run were also heterogeneous in peak density, as their interquartile ranges demonstrate.