A Hypergeometric Probability Model for Protein Identification and Validation Using Tandem Mass Spectral Data and Protein Sequence Databases

ARTICLE *in* ANALYTICAL CHEMISTRY · SEPTEMBER 2003

Impact Factor: 5.64 · DOI: 10.1021/ac034157w · Source: PubMed

Rovshan G Sadygov
University of Texas Medical Branch at Galves…
51 PUBLICATIONS 3,815 CITATIONS

John R Yates
The Scripps Research Institute
616 PUBLICATIONS 60,394 CITATIONS

A Hypergeometric Probability Model for Protein Identification and Validation Using Tandem Mass Spectral Data and Protein Sequence Databases

Rovshan G. Sadygov and John R. Yates, III*

Department of Cell Biology, SR11, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037

We present a new probability-based method for protein identification using tandem mass spectra and protein databases. The method employs a hypergeometric distribution to model frequencies of matches between fragment ions predicted for peptide sequences with a specific $(M + H)^+$ value (at some mass tolerance) in a protein sequence database and an experimental tandem mass spectrum. The hypergeometric distribution constitutes null hypothesis—all peptide matches to a tandem mass spectrum are random. It is used to generate a score characterizing the randomness of a database sequence match to an experimental tandem mass spectrum and to determine the level of significance of the null hypothesis. For each tandem mass spectrum and database search, a peptide is identified that has the least probability of being a random match to the spectrum and the corresponding level of significance of the null hypothesis is determined. To check the validity of the hypergeometric model in describing fragment ion matches, we used χ^2 test. The distribution of frequencies and corresponding hypergeometric probabilities are generated for each tandem mass spectrum. No proteolytic cleavage specificity is used to create the peptide sequences from the database. We do not use any empirical probabilities in this method. The scores generated by the hypergeometric model do not have a significant molecular weight bias and are reasonably independent of database size. The approach has been implemented in a database search algorithm, PEP_PROBE. By using a large set of tandem mass spectra derived from a set of peptides created by digestion of a collection of known proteins using four different proteases, a false positive rate of 5% is demonstrated.

Tandem mass spectrometry (MS/MS) in combination with database searching is an important tool for proteomics. Protein samples are enzymatically digested, and the resulting peptides are analyzed by tandem mass spectrometry.[1,2] The fragmentation information in a tandem mass spectrum of a peptide can be used to search against a protein sequence database to identify the amino acid sequence represented in the spectrum. This method has a number of practical uses from identifying proteins separated by 2-DGE to identifying proteins in complexes, organelles, whole-cell lysates, and tissue homogenates.[3]

A common element of all database-searching algorithms is the assignment of a score that signifies the quality of the match between the sequence and spectrum. The score is a measure of the closeness of fit between the spectrum and the peptide sequence retrieved from the database. Quantitative measures such as cross-correlation, dot product, and fragment ion frequency-based probabilities have been used to measure the closeness of fit between spectrum and sequence. Several algorithms have been developed to search databases using these methods.[4−11] An emerging issue is the significance of the match between a peptide sequence and a tandem mass spectrum.[5,12] A database-searching program will always find a sequence correlating with the tandem mass spectrum. However, not all matches are statistically significant, with some matches occurring at random and others occurring between poor-quality tandem mass spectra, spectra exhibiting odd fragmentation patterns, or spectra with unanticipated modifications. It is important, therefore, that a scoring scheme provide a means of estimating the statistical significance of each peptide match.

Most of the database-searching algorithms reported in the literature use a database sequence and the experimental tandem mass spectrum to determine the match score. SEQUEST determines a score by calculating a cross-correlation between the experimental and 500 theoretical spectra reconstructed from sequences with the fit closest to the tandem mass spectrum.[4] The cross-correlation function provides a quantitative measure of the closeness of fit to the sequence and the relative closeness of fit of other sequences, but the method does not a priori provide the statistical significance of the match.[13,14] A normalization method

* Corresponding author: (tel) (858)784-8862; (fax) (858) 784-8883; (e-mail) jyates@scripps.edu.

[1] Yates, J. R., III, *Electrophoresis* **1998**, *19*, 893−900.

[2] Aebersold, R.; Goodlett, D. R. *Chem. Rev.* **2001**, *101*, 269−295.

[3] Florens, L.; et al. *Nature* **2002**, *419*, 520−526.

[4] Eng, J. K.; McCormack, A. L.; Yates, J. R., III. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−989.

[5] Perkins, D. N.; et al. *Electrophoresis* **1999**, *20*, 3551−3567.

[6] Zhang, W.; Chait, B. T. *Anal, Chem*, **2000**, *72*, 2482−2489.

[7] Zhang, N.; Aebersold, R.; Schwikowski, B. *Proteomics* **2002**, *2*, 1406−1412.

[8] Bafna, V.; Edwards, N. *Bioinformatics* **2001**, *17*, S13−S21.

[9] Hansen, B. T.; Jones, J. A.; Mason, D. E.; Liebler, D. C. *Anal. Chem.* **2001**, *73* (8), 1676−1683.

[10] Fenyo, D.; Beavis, R. C. *Anal. Chem.*, in press.

[11] Havilo, M.; Haddad, Y.; Smilansky, Z. *Anal. Chem.* **2003**, *75* (3), 435−444.

[12] Eriksson, J.; Chait, B. T.; Fanyo, D. *Anal. Chem.* **2000**, *72* (5), 999−1005.

10.1021/ac034157w CCC: $25.00 © 2003 American Chemical Society
Published on Web 06/21/2003

has been proposed to eliminate the molecular weight dependency of the cross-correlation method to allow an empirical statistical significance of the match to be generated.[14] PROFOUND uses a Bayesian model to calculate the probability that the candidate peptide is a true match.[6] This model comprises three spaces—background, experimental data, and database. The main term of the final probability is the conditional probability that the experimental data are produced by the theoretical peptide and the background information. The same model has been used in ProbID[7] with the addition of new empirical terms. Neither sequence tagging[15] nor pattern recognition[9] methods use collective properties of the database to identify sequences; i.e., peptide scores do not depend on any overall characteristics of a database, such as the number fragment ions or the number of them that match the experimental spectrum.

To our knowledge, only MOWSE[16] and MASCOT [5] (these programs employ similar probability models) use accumulated information generated from the database to determine match scores. To calculate scores, MOWSE uses the normalized frequencies of peptide $(M + H)^+$ distributions from the database. However, the validity of the statistical model used in MOWSE is not easily verified. Even though MOWSE uses a probability model, it is not clear if any single probability space can be generated for the model. The model consists of numerous spaces separated by the mass of an average amino acid. The probabilities that are used are normalized within each window but not to the overall database. It is presumed a similar model is used for MASCOT MS/MS database searches, but no details of this model have been published.

By using the overall properties of peptide $(M + H)^+$ values in a database in relation to the experimental tandem mass spectrum, the quality of a sequence match relative to the background of random sequence matches can be measured. In this paper, we present a new probabilistic model based on the amino acid sequence properties of a database. We show that the frequency of fragment ion matches to a tandem mass spectrum follows a hypergeometric distribution. The distribution predicts a probability of generating any given number of fragment ion matches to an experimental tandem mass spectrum at random. To test the hypothesis that fragment ion match frequency and hypergeometric probabilities belong to the same distribution, we use the χ^2 test. The hypergeometric distribution constitutes the null hypothesis, H_0, which postulates that any database peptide match to tandem mass spectrum is random. A match score for a peptide is presented as $-\log(P)$, where P is the hypergeometric probability. The peptide that has the least probability of being a random match is identified as the best sequence fit for the tandem mass spectrum. The significance of the peptide match is calculated from the level of significance at which H_0 can be rejected.

Unlike some of the recently introduced probabilistic models (ProbID[7]) our model does not employ any empirical parameters. Use of empirical parameters may introduce biases for specific experiments, instruments or fragmentation mechanisms. It also does not require a priori computation of fragment ions in a database (MOWSE). Our model depends only on the sequences present in the database. Since the scores are dependent on the frequency distribution and the latter is determined by the databases, the scores generated by the hypergeometric model are in principle database dependent. But as we show below, the actual dependency is negligible. The approach has been implemented in a database search program called PEP_PROBE. The program is written in java and can be run on a single CPU or multiple CPUs operated in parallel.

EXPERIMENTAL SECTION

Standard Preparation and Digestion. A mixture of proteins was used as a standard to validate database searching and to study the algorithm. The mixture of proteins was treated in the following manner. A mixture of proteins containing equimolar levels of phosphorylase a (rabbit skeletal muscle), cytochrome *c* (horse), apomyoglobin (horse heart), albumin (bovine serum), and β-casein (bovine) was used for all experiments. The resulting mixture (~1 nmol/μL in water) was adjusted to 8 M urea with the addition of solid urea, reduced with dithiothreitol (20 mM final concentration at 50 °C for 20 min), and alkylated with iodoacetamide (50 mM final concentration in the dark at room temperature). The denatured, reduced, and alkylated protein mixture was divided into four aliquots, and each was digested using a different protease. Aliquot 1 was diluted 3-fold with 100 mM Tris, pH 8.5, and $CaCl_2$ was added to a final concentration of 1 mM. Modified trypsin (Promega) was added at a 1:100 enzyme-to-substrate ratio (w/w), and the mixture was incubated overnight at 37 °C with constant mixing (Thermomixer, Eppendorf). Aliquot 2 was diluted 3-fold with 100 mM Tris, pH 8.5, elastase (Roche) was added at a 1:50 enzyme-to-substrate ratio (w/w), and the resultant mixture was incubated overnight with mixing at 37 °C. Aliquot 3 was diluted 3× with 100 mM Tris, pH 8.5, subtilisin (Sigma) was added at a 1:50 enzyme-to-substrate ratio (w/w), and the resultant mixture was incubated with mixing for 3 h at 37 °C. Aliquot 4 was adjusted to pH 11 with 1 M NaOH, proteinase K (Roche) was added at a 1:100 enzyme-to-substrate ratio (w/w), and the resultant mixture was incubated at 37 °C for 3 h with constant mixing. Each digestion was quenched with the addition of formic acid to 5% and frozen at −80 °C until analysis by MudPIT as described below.

Multidimensional Protein Identification Technology. Peptide mixtures were analyzed as previously described in MacCoss et al.[17] A triphasic microcapillary column was constructed from 100-μm-i.d. fused-silica capillary tubing pulled to a 5-μm-i.d. tip using a Sutter Instruments P-2000 CO_2 laser puller (Novato, CA). Each column was packed with 7 cm of 5-μm Aqua C18 material (Phenomenex, Ventura, CA) and 3 cm of 5-μm Partisphere strong cation exchanger (Whatman, Clifton, NJ), followed by another 3 cm of Aqua C18. The columns were equilibrated with 5% acetonitrile/0.1% formic acid, and ~4 pmol of each protein digest was loaded directly onto separate capillary columns using a high-pressure bomb.

After loading the peptide digests, the column was placed in-line with a Surveyor quaternary HPLC (ThermoFinnigan, Palo Alto, CA) and analyzed using a modified six-step separation

[13] Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74* (20), 5383−5392.

[14] MacCoss, M. J.; Wu, C. C.; Yates, J. R., III. *Anal. Chem.* **2002**, *74,* 5593−5599.

[15] Mann, M.; Wilm, M. *Anal. Chem.* **1994**, *66*, 4390−4399.

[16] Pappin, D. J. C.; Hojrup, P.; Bleasby, A. J. *Curr. Biol.* **1993**, *3*, 327−332.

[17] MacCoss, M. J.; et al. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 7900−7905.

described previously.[17] The buffer solutions used were 5% acetonitrile/0.1% formic acid (buffer A), 80% acetonitrile/0.1%formic acid (buffer B), and 500 mM ammonium acetate/5% acetonitrile/0.1% formic acid (buffer C). Step 1 consisted of a 100-min gradient from 0 to 100% buffer B. Steps 2−5 had the following profile: 3 min of 100% buffer A, 2 min of *X*% buffer C, a 10-min gradient from 0 to 15% buffer B, and a 97-min gradient from 15 to 45% buffer B. The 2-min buffer C percentages (*X*) were 10, 20, 30, 40, and 50% for the six-step analysis. For the final step, the gradient contained the following: 3 min of 100% buffer A, 20 min of 100% buffer C, a 10-min gradient from 0 to 15% buffer B, and a 107-min gradient from 15 to 70% buffer B.

As peptides eluted from the microcapillary column, they were electrosprayed directly into an LCQ-Deca mass spectrometer (ThermoFinnigan, Palo Alto, CA) with the application of a distal 2.4-kV spray voltage. A cycle of one full-scan mass spectrum (400−1400 *m*/*z*) followed by three data-dependent tandem mass spectra at a 35% normalized collision energy was repeated continuously throughout each step of the multidimensional separation. The application of all mass spectrometer scan functions and HPLC solvent gradients was controlled by the Xcaliber data system.

Database Searching. Tandem mass spectra were extracted from the Xcaliber data files using the program ExtractMS and written to a separate file. ExtractMS assigns charge states to +1 spectra and then calculates the $(M + H)^+$ values for all other spectra for both a +2 and +3 charge states. This file is then processed by the 2to3 program to reassign charge states to all +2 and +3 charge-state tandem mass spectra as one or the other and to remove poor-quality tandem mass spectra.[18] This data file was then used by PEP_PROBE to search databases. Peptides from the digested protein standard were searched through the nonredundant protein sequence database with 907 646 proteins obtained from the National Center for Biotechnology Information (downloaded on 04/15/2002). The yeast proteasome was searched through a yeast protein sequence database. Mass tolerance for $(M + H)^+$ values used in the search was 3.0 amu, and fragment ion mass tolerance was 0.4 amu. No enzymatic cleavage specificity was used to select amino acid sequences during the search process.

THEORY

The basis of the approach is to formulate the distribution of frequencies for fragment ion matches in a sequence database as a probabilistic model. Let us designate the number of all predicted fragment ions in a sequence database from sequences matching a specific $(M + H)^+$ value as *N* and the number of all these fragment ions matching a peak in an experimental tandem mass spectrum as *K*. The sample space has a total of *N* elements and *K* of them match peaks in the tandem mass spectrum ($K \leq N$). If a peptide has *L* amino acids and K_1 of its fragment ions match to the experimental tandem mass spectrum, then we seek the hypergeometric probability that the event is random, which consists of K_1 matches in N_1=2(L − 1) tries (we consider only two types of fragment ions, b- and y-ions). The corresponding probability is given by

$$P_{K,N}(K_1,N_1) = \frac{C_K^{K_1} C_{N-K}^{N_1-K_1}}{C_N^{N_1}}$$

where C_I^J is the binomial coefficient. The calculated hypergeometric distribution accounts for the matches from all database sequences. It is database and spectrum dependent and independent of protease specificity.

The model can be extended to include other ion types and fragmentation models. However, it should be noted that the model (as applied here) treats all ion types equally. Therefore, if several fragmentation patterns are encoded, the results may exaggerate dissociation models of marginal experimental importance. This trend, of course, would be true for all scoring schemes.

We assume that all matches of fragment ions (consequently of peptides from which the ions are derived) to a tandem mass spectrum are random. This is the null hypothesis, H_0, and the distribution of predicted fragment ions in a database is approximated by a hypergeometric distribution. The hypergeometric distribution provides the probability that a peptide match to the experimental tandem mass spectrum occurs at random. This probability is used to generate a score for a match between experimental spectrum and database peptide. The smaller the probability of a random match, the more likely a match is a true positive. The alternative to the null hypothesis, H_1, is that a peptide match to a tandem mass spectrum is not at random but is a true match. Therefore, the significance of the match for a peptide is determined from the level of the significance at which the null hypothesis can be rejected for this peptide.

This approach and statement of the problem can be generalized to enable us to incorporate other probabilistic models as well. In what follows, we show how one can state the problem in terms of Poisson's model. Given the number of all fragment ion matches, *K*, and number of all fragment ions, *N*, the probability of a match per fragment ion is

$$p = K/N$$

Then Poisson's probability that K_1 matches will occur by chance in N_1 tries is given by

$$P(K_1,p) = \frac{(N_1p)^{K_1}}{K_1!} \exp(-N_1p)$$

This model also allows us to calculate the probability that the match between a database sequence and a spectrum is a chance event. There is a difference between the hypergeometric and the Poisson models. The Poisson model assumes sampling with a replacement, while the hypergeometric model is based on sampling without a replacement. Sampling with a replacement assumes that for each sampling the probability of a match is the same. In the example of matching fragment ions in a tandem mass spectrum to those predicted in a sequence database, the probabilities of fragment ion matches would be different. Therefore sampling without a replacement is a more appropriate assumption for searches of databases using tandem mass spectra. Both of the models have been implemented in the PEP_PROBE program. PEP_PROBE can be used to search a database using hypergeometric or Poisson models independently or in combination.

[18] Sadygov, R.; Eng, J.; Durr, E.; Saraf, A.; McDonald, H.; MacCoss, M. J.; Yates, J. R., III. *J. Proteome Res.* **2002**, *1* (3), 211−215.

The hypergeometric distribution approximates the frequencies of fragment ion matches (assumed to be random) of peptide sequences in a database to an experimental spectrum. To verify the validity of the approximation, we used a χ^2 test. The null hypothesis is that the frequency of fragment ion matches and hypergeometric probabilities belong to the same distribution. Statistics are constructed in the following way. The number of fragment ion matches constitutes the bins and the number of peptide sequences in a database (whose mass is equal to the precursor peptide's mass with some accuracy) having that many matches constitutes the observations. The χ^2 test is

$$\chi^2 = \sum_{i=0}^{2(L-1)} \frac{(K_i - k_i)^2}{k_i}$$

where K_i and k_i are the observed and expected number of peptides having i matches, respectively. The χ^2 distribution is given by the χ^2 probability function $Q(\chi^2|F)$, which is an incomplete γ function. $Q(\chi^2|F)$ is interpreted as the probability that the sum of squares of $2(L - 1)$ random normal variables of unit variance will be greater than χ^2. The test is incorporated into the PEP_PROBE program, performed for each spectrum, and the probability of rejecting the null hypothesis is printed in the output.

We implemented two methods in PEP_PROBE to calculate peptide match significance. The first method uses the so-called P-value.[19] For this we calculate the H_0 (the match is random) probability of obtaining the observed value of test statistic (number of matching fragment ions) or more extreme in the direction indicated by the alternative hypothesis, H_1 (the match is a true positive). The P-value is calculated from a cumulative hypergeometric distribution. If P_0 is the probability for the peptide identification at random, then the expectation value for database peptide sequences in a database that will be less random is

$$E = M\Phi(P < P_0) = M \sum_{P > P_0} P$$

where Φ is the cumulative distribution function given by the hypergeometric model, M is the number of all peptides in a database matching an $(\mathrm{M} + \mathrm{H})^+$ value. Interpretation of the expectation value is how many peptides from the database are expected to have the same or better matches to the experimental spectrum. Thus, a value of 0.1 means that if the database size is increased 10 times, then we expect one peptide with the same or better quality match as the peptide under consideration to appear by chance. Normally peptides with expectation values of 0.05 or smaller are considered as significant hits.

The second method for peptide identification implemented in PEP_PROBE employs Chebyshev's inequality to calculate the upper bound of the probability that a random match can be expected to have less probability than the identified peptide. The computation uses information about the average number of fragment ion matches per peptide and its variation. Chebyshev's inequality does not assume any specific probability distribution.

Two kinds of errors are encountered when peptide sequences from a database are matched to tandem mass spectra. The first kind of error is that the null hypothesis (peptide match is random), H_0, is correct but it has been rejected. This situation results in accepting the wrong peptide match or a false positive. The second type of error occurs when null hypothesis is accepted but the alternative (peptide match is real), H_1, is true. P-value and Chebyshev's inequality help us to reduce the significance of the first kind of error.

RESULTS AND DISCUSSION

PEP_PROBE matches a tandem mass spectrum to a sequence that is the least likely to be a random match based on the hypergeometric distribution of predicted fragment ions in a protein sequence database. The corresponding score is the probability that the peptide sequence is a random match to the tandem mass spectrum assuming that the distribution of frequencies of fragment ion matches is hypergeometric. For a correct match between a spectrum and a sequence, the probabilities are small values. Therefore, probabilities are expressed as a minus base 10 logarithm, $-\log(P)$. After this conversion, the peptide with lowest probability of being A random match has the highest score. Interpretation of scores is simple, since the minus of the score to the power of 10 is the probability that the match is random.

The experimental tandem mass spectrum shown in Figure 1 was searched against the *Saccharomyces cerevisiae* database, and PEP_PROBE identified the peptide sequence ATHILDFGPGGASGLGVLTHR as having the lowest probability of being a random match. This sequence corresponds to the correct answer, and the score is 26.62 (e.g., probability of random match is $10^{-26.62}$). The overall number of fragment ions generated from the database is 5 284 150, and 569 160 of them have matches to the spectrum. The peptide has 40 fragment ions, and 34 of them matched to the spectrum. These data are used in the above formula to generate the hypergeometric probability score for the peptide.

The second best match to the tandem mass spectrum is LTPPQLPPQLENVILNKY. Its probability of random match is $10^{-6.19}$. Fifteen fragment ions of this peptide matched to the experimental spectrum. Poisson's probabilities for the two highest ranking peptides are $10^{-18.77}$ and $10^{-5.25}$, respectively.

The theoretical population (based on hypergeometric distribution) and observed fragment ion frequencies (normalized) versus number of matches are depicted in Figure 2. As is evident from the figure, the predicted number of matches from the model distribution is in excellent agreement with the experimental observation. This is confirmed by the χ^2 test, which determined that the probability the data sets came from different distributions is equal to 0.

The match score is generated from the hypergeometric probability distribution, which is the model for database fragment ion distribution. The expectation is that frequency distributions may be database dependent; therefore, it is important to see how much the distribution changes with the number of protein sequences in the database. Figure 3 shows the theoretical distributions (based on the hypergeometric model) for the *S. cerevisiae* (6200 sequences) and NRP (907 646 sequences) sequence databases. The distributions are not identical but very close. The score from NRP database search is 26.51 (the random probability is $10^{-26.51}$). The scores (26.62 and 26.51) generated

[19] Ewens, W. J.; Grant, G. R. *Statistical Methods in Bioinformatics. Statistics for Biology and Health*; Springer-Verlag: New York, 2002.

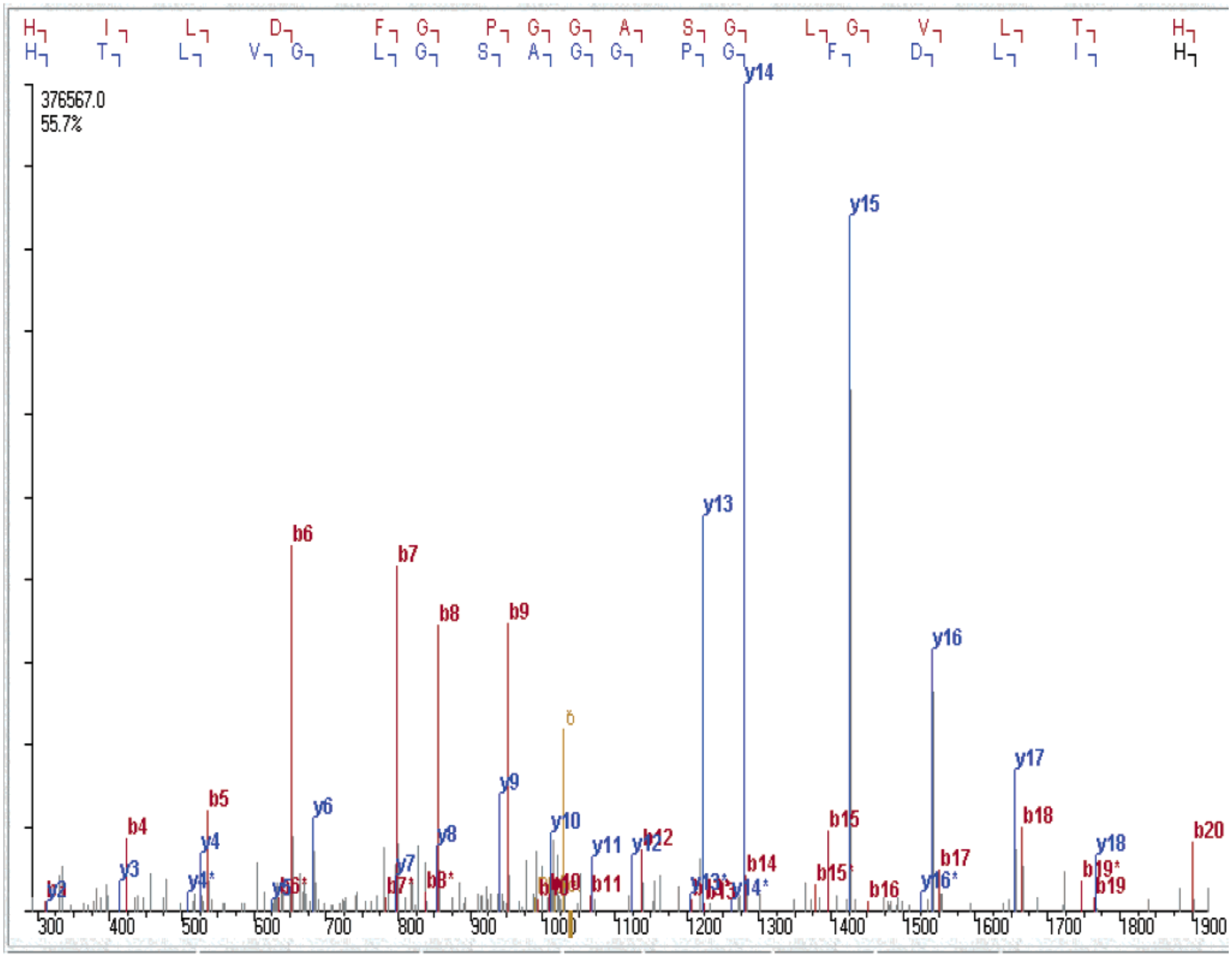

Figure 1. Tandem mass spectrum of the peptide ATHILDFGPGGASGLGVLTHR. The page has been generated by the DTASelect program.[22] This peptide is derived from the yeast protein FAS1, which plays a role in fatty acid synthesis. Sequence ions of types b and y are shown in the tandem mass spectrum.

for searches of the two different databases are only slightly different despite the large difference in size of the two databases.

There are two hypotheses in our approach. The null hypothesis is that a peptide match to a tandem mass spectrum is random (i.e., it is a false positive). The alternative hypothesis is that the match is a true positive. The distribution of the null hypothesis is approximated by hypergeometric probabilities. For each number of fragment ion matches and peptide $(M + H)^+$ values, the hypergeometric distribution provides the probability that the peptide matched to the spectrum at random. Thus, a peptide that has the least probability of being a random match is the most likely candidate for identification. However, not every identified peptide is statistically important. A peptide may be the least random match to a tandem mass spectrum, but it is not necessarily a true positive. To determine statistical significance of a peptide match to a tandem mass spectrum, we use a technique that allows us to specify the level of significance of the peptide match to the null hypothesis. This is done using a *P*-value. *P*-value is the expected number of peptides from the database that would have better matches (to the spectrum) than the candidate peptide. The *P*-value of a match between a sequence and a tandem mass spectrum can be calculated from the cumulative distribution of the model probabilities (the hypergeometric in this case).[19] The cumulative distribution function for the peptide ATHILDFGPGGASGLGVLTH and its tandem mass spectrum (Figure 1) is presented in Figure 4 (result is from S. *cerevisiae* database). The *P*-value of the search is zero. The confidence of the match calculated from the Chebyshev's inequality is 97.4%. The *P*-value for the second best matching peptide, LTPPQLPPQLENVILNKY, is 19 and the confidence from the Chebyshev's inequality is 88.9%.

Database searching programs that generate closeness of fit measures that are sums or products of fragment ion matches in the spectrum can artificially inflate as a function of molecular weight. Larger peptides have more fragment ions, and unless a correction for molecular weight or number of fragment ions is made, tandem mass spectra representing larger peptides often have higher scores. Peptides with higher charge states, e.g., +3, tend to be of higher molecular weight and, thus, show a trend toward higher scores (multiply charged fragment ions are often taken into account as well). For example, scores generated by SEQUEST's cross-correlation algorithm scale with molecular weight with triply charged tandem mass spectra producing higher cross-correlation scores. This trend hinders the use of simple and uniform criteria for large-scale peptide identification. MacCoss et

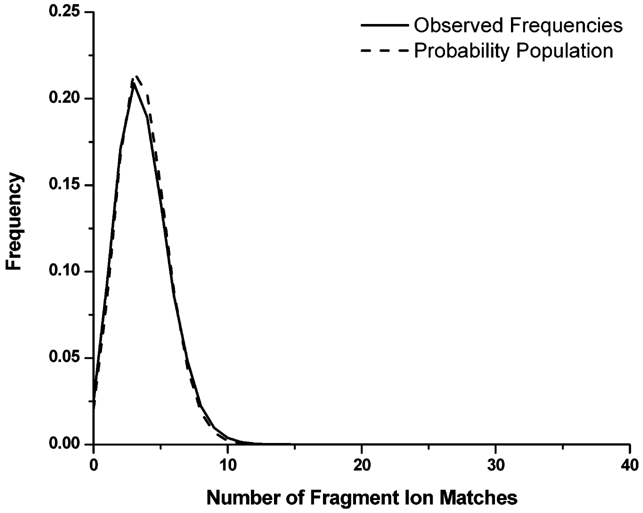

Figure 2. Normalized frequencies of peptide fragment ion matches vs number of matches for the tandem mass spectrum of the peptide ATHILDFGPGGASGLGVLTHR (solid line). Overlaid on the graph for the actual distribution is the theoretical distribution predicted by a hypergeometric distribution (dashed line).

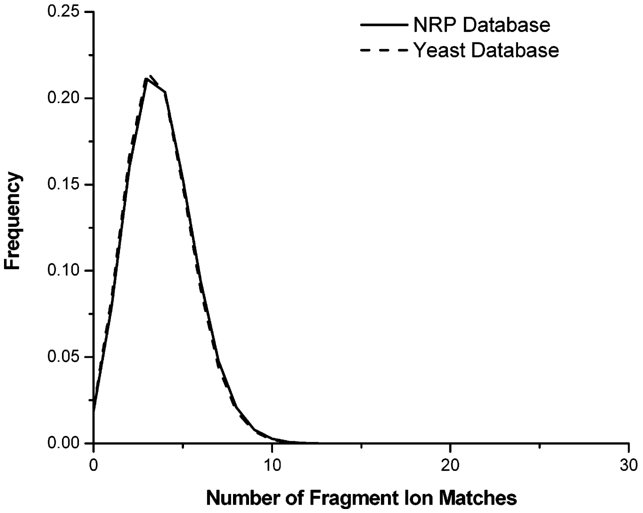

Figure 3. Normalized frequencies of fragment ion matches for the peptide ATHILDFGPGGASGLGVLTHR obtained from the *S. cerevisiae* database, 6200 proteins (dashed line), and the Non Redundant Protein Sequence database, 907 646 sequences (solid line).

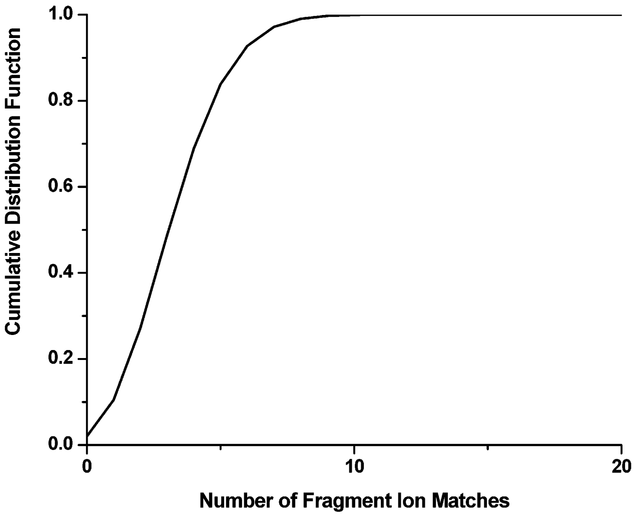

Figure 4. Cumulative distribution function of hypergeometric probabilities for the spectrum and peptide from Figure 1.

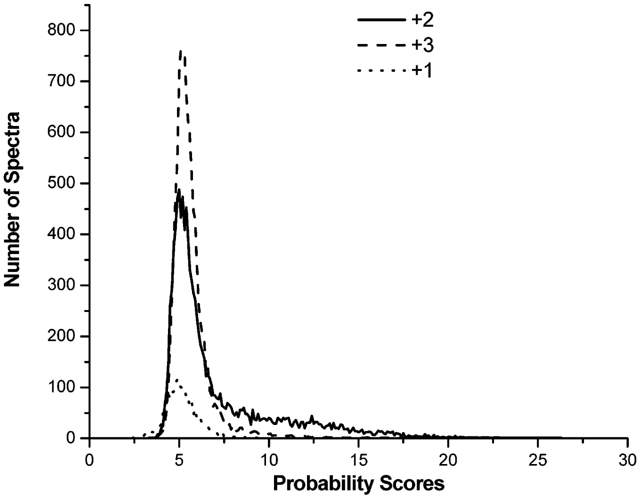

Figure 5. Distribution of hypergeometric probability scores by precursor peptide charge states for a data set of tandem mass spectra obtained from the digested yeast proteasome and searched against the *S. cerevisiae* sequence database.[20] Dotted line is the curve for charge state +1, solid line is the curve for charge state +2, and dashed line is the curve for the charge state +3.

al.[14] suggested a method to normalize SEQUEST scores to provide more uniform criteria for assessment of search scores and determination of an empirical statistical confidence. To determine whether the hypergeometric probability model shows a bias toward higher charge state (e.g., molecular weight), the distribution of probability scores as a function of charge state was plotted as shown in Figure 5. The scores obtained from the probability model do not show charge-state dependence for +2 and +3 charges. The data set contained ~23 000 tandem mass spectra generated from the proteasome isolated from *S. cerevisiae*.[20]

The accuracy of the program was tested using a set of tandem mass spectra derived from the MudPIT analysis of a digested protein mixture. This data set was used previously to validate the use of normalized XCorr values. The data set contains ~59 000 tandem mass spectra obtained from a known mixture of peptides.[14] Figure 6 plots the distribution of probability scores for all spectra in the data set when searched against the nonredundant database. The peptide identifications are determined to be true (solid line) if the identified peptide belongs to one of the proteins known to be in the mixture, and false (dashed line) otherwise. Out of the 59 000 tandem mass spectra in the data set, 5000 are true identifications. Of these, only about 300 are +1 charged tandem mass spectra. The first shoulder in the false identification plot is due to singly charged tandem mass spectra. This is seen from Figure 7 where we plot the same data, but separately for +1 and combined for +2 and +3 charge states. False identifications of +1 charge-state tandem mass spectra form a separate peak at lower probability than those of +2 and +3 charge-state spectra. The true identifications of tandem mass spectra of singly charged peptides are too few to plot (~300), as they make up 6% of the total true identifications. The tandem mass spectra produced in

[20] Verma, R.; et al. *Mol. Cell* **2001**, *8*, 439−448.

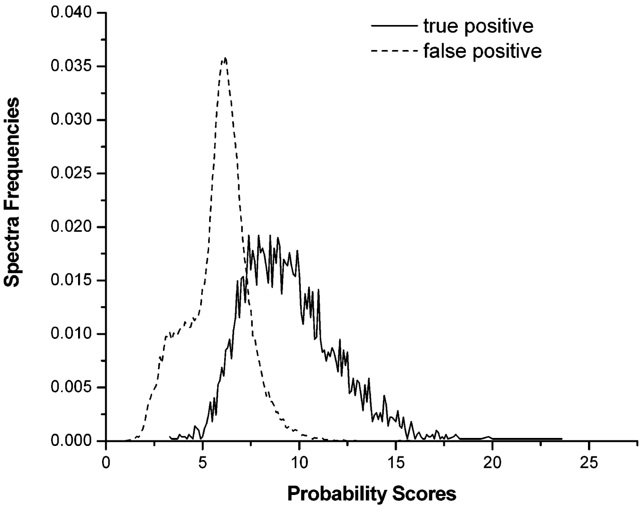

Figure 6. Distribution of probability scores of true (solid line) and false (dashed line) positives obtained by searching a mixture of known peptide tandem mass spectra against NRP database (907 646 sequences).

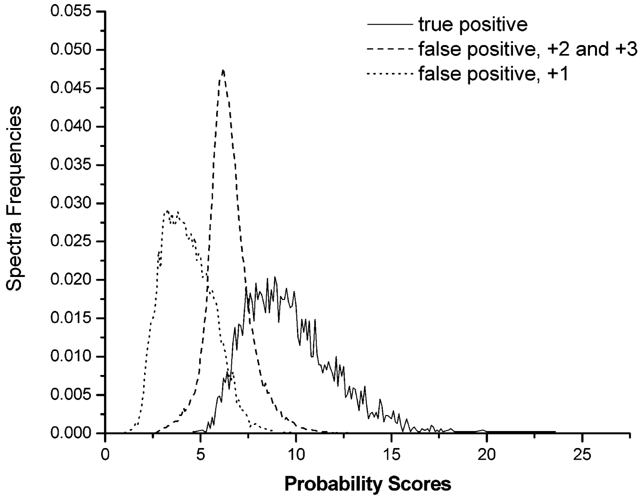

Figure 7. Distribution of probability scores plotted for +1 charge-state tandem mass spectra representing false positives (dashed−dotted line). The +2 and +3 charge-state tandem mass spectra are plotted showing false positives (dashed line) and true positives (solid line). True positives of +1 charge-state tandem mass spectra are not shown because of the small number of spectra.

an ion trap mass spectrometer of singly charged peptides generally produce spectra with limited sequence information, and they are often problematic for database searching.

Table 1 shows the percentage of the false identifications by the probability score. At a score of 8 half of the identifications are false, while at a score of 12 and above at most 5% of the identifications are false. These results are in a very good agreement with what we calculate for the significance of matches. By identifying appropriate values for the statistical significance of matches, the evaluation of results for large-scale studies should be improved.

Table 1. Percentage of False Identifications for Several Selected Probability Scores[a]

probability scores	% of false identification
8	49.73
9	29.98
10	16.4
11	9.35
12	5.38
13	4.03

[a] For example, 49.7% of spectra that scored 8 or higher are false identifications, while only 5.38% of spectra that scored 12 or higher are false identifications (See the text for the criteria that were used to determine false and true identifications.).

CONCLUSION

We have developed a probabilistic model based on a hypergeometric distribution to identify peptides from their respective tandem mass spectra and a protein sequence database. The null hypothesis is that all matches of fragment ions derived from a sequence database to a tandem mass spectrum are at random. The frequencies of these fragment ion matches to a tandem mass spectrum are modeled by hypergeometric distribution. From the distribution, we calculate the probability that a peptide sequence match to the experimental spectrum is random. This probability is reported as an identification score. The significance of peptide match is determined from the level of significance at which the null hypothesis can be rejected. We show how the statistical significance of the peptide match is computed from the cumulative distribution function of the frequencies. The calculated scores are in principle database dependent, but the dependency is shown to be very small. The scores do not show a bias toward tandem mass spectra of higher charge states, and the approach can be extended to other statistical models, like a Poisson distribution, when appropriate.

The probability model employed does not use information about ion abundance. Fragment ion abundance can be an indication of the relative importance of ions in a tandem mass spectrum. In general, sequence ions have greater ion abundance in a tandem mass spectrum than the neutral losses associated with those fragment ions. Furthermore, Tabb et al. have shown that there is a statistical relationship between fragment ion abundance and the type of amino acid residue at the position in the sequence.[21]This trend may be useful to increase the specificity of matches.

Matching sequences to tandem mass spectra using frequency distributions of predicted fragment ions as a measure for random matches may under represent correct matches. In future work, we will incorporate fragment ion abundance information into the probability model presented here.

ACKNOWLEDGMENT

The authors thank Hayes McDonald, Mike MacCoss, and John Venable for fruitful discussions and Hayes McDonald, Mike MacCoss, and Christine Wu for making available the tandem mass spectra used in this study.

Received for review February 17, 2003. Accepted May 12, 2003.

AC034157W

(21) Tabb, D. L.; Smith, L. L.; Breci,L. A.; Wysocki, V. H.; Lin, D.; Yates, J. R., III. *Anal. Chem.* **2003**, *75*, 1155−1163

(22) Tabb, D. L.; McDonakd, W. H.; Yates, J. R., III. *J. Proteome Res.* **2002**, *1* (1), 21−26.