Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search

Andrew Keller,*[,†] Alexey I. Nesvizhskii,*[,†] Eugene Kolker, and Ruedi Aebersold

Institute for Systems Biology, 1441 North 34th Street, Seattle, Washington 98103

We present a statistical model to estimate the accuracy of peptide assignments to tandem mass (MS/MS) spectra made by database search applications such as SEQUEST. Employing the expectation maximization algorithm, the analysis learns to distinguish correct from incorrect database search results, computing probabilities that peptide assignments to spectra are correct based upon database search scores and the number of tryptic termini of peptides. Using SEQUEST search results for spectra generated from a sample of known protein components, we demonstrate that the computed probabilities are accurate and have high power to discriminate between correctly and incorrectly assigned peptides. This analysis makes it possible to filter large volumes of MS/MS database search results with predictable false identification error rates and can serve as a common standard by which the results of different research groups are compared.

A major goal of proteomics research is to catalog and quantify the proteins and protein complexes present in cells grown under a variety of conditions.[1,2] Tandem mass spectrometry (MS/MS) has been particularly useful for determining the protein components of complex mixtures.[3−6] Proteins in a sample are first digested into smaller peptides, usually by the enzyme trypsin, and subjected to reverse-phase chromatography. Peptides are then ionized and fragmented to produce signature MS/MS spectra that are used for identification. Most frequently, peptide identifications are made by searching MS/MS spectra against a sequence database to find the best matching database peptide.[7] From these peptide assignments to spectra, the original proteins present in the sample are inferred. Posttranslational modifications to peptides can be investigated by searching spectra against a database while allowing for specific peptide modifications.[8−10] Labeling methods, such as differential isotopic labeling of cysteines with the ICAT reagent, can be combined with MS/MS database search to quantify the levels of proteins in one sample relative to those in a reference.[11−13] In addition, peptides corresponding to MS/MS spectra can be derived without a database search by de novo sequencing.[13−15] This is particularly useful for samples from organisms with polymorphic mutations or unsequenced genomes. De novo sequencing can also be combined with a database search.[14,16−17]

Over the past few years, MS/MS with database search has been used increasingly for high-throughput analysis of complex protein samples. This has been made possible by automated database search software such as SEQUEST,[18] Mascot,[19] and Sonar.[20] These applications compare each spectrum against those expected for all possible peptides obtained from a sequence database that have masses within an error tolerance of the precursor ion mass. Each spectrum is then assigned the database peptide with the highest overall score, or set of scores, that reflects various aspects of the fit between spectrum and peptide. These scores help discriminate between correct and incorrect peptide assignments to spectra and, hence, facilitate detection of false identifications. They are often based on counts of common

* Corresponding authors. Fax: 206-732-1299. E-mails: akeller@systemsbiology.org; nesvi@systemsbiology.org.

[†] These authors contributed equally to this work.

(1) Pandey, A.; Mann, M. *Nature* **2000**, *405*, 837−846.

(2) Aebersold, R.; Goodlett, D. R. *Chem. Rev.* **2001**, *101*, 269−287.

(3) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; et al. *Nat. Biotechnol.* **1999**, *17*, 676−682.

(4) Washburn, M. P.; Wolters, D.; Yates, J. R. *Nat. Biotechnol.* **2001**, *19*, 242−247.

(5) Gavin, A.; Bosche, M.; Krause, R.; Grandi, P.; et al. *Nature* **2002**, *415*, 141−147.

(6) Ho, Y.; Gruhler, A.; Heilbut, A.; Bader, G. D.; et al. *Nature* **2002**, *415*, 180−183.

(7) Fenyo, D. *Curr. Opin. Biotechnol.* **2000**, *11*, 391−395.

(8) Yates, J. R., III; Eng, J. K.; McCormack, A. L.; Schieltz, D. *Anal. Chem.* **1995**, *67* (8), 1426−1436.

(9) Pevzner, P. A.; Mulyukov, Z.; Dancik, V.; Tang, C. L. *Genome Res.* **2001**, *11*, 290−299.

(10) Liebler, D. C.; Hansen, B. T.; Davey, S. W.; Tiscareno, L.; Mason, D. E. *Anal. Chem.* **2002**, *74*, 203-210.

(11) Oda, Y.; Huang, K.; Cross, F. R.; Cowburn, D.; Chait, B. T. *Proc. Natl. Acad. Sci.* **1999**, *96*, 6591−6596.

(12) Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R. *Nat. Biotechnol.* **1999**, *17*, 994−999.

(13) Goodlett, D. R.; Keller, A.; Watts, J. D.; Newitt, R.; et al. *Rapid Commun. Mass Spectrom.* **2001**, *15*, 1214−1221.

(14) Taylor, J. A.; Johnson, R. S. *Rapid Commun. Mass Spectrom.* **1997**, *11*, 1067−1075.

(15) Chen, T.; Kao, M.; Tepel, M.; Rush, J.; Church, G. *The 11th Annual SIAM-ACM Symposium on Discrete Algorithms (SODA 2000)*, San Francisco, 2000; ACM Press: New York; pp 389−398.

(16) Mann, M.; Wilm, M. *Anal. Chem.* **1994**, *66* (24), 4390−4399.

(17) Shevchenko, A.; Sunyaev, S.; Loboda, A.; Shevchenko, A.; et al. *Anal. Chem.* **2001**, *73*, 1917−1926.

(18) Eng, J. K.; McCormack, A. L.; Yates J. R. III. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−989.

(19) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551−3567.

(20) Field, H. I.; Fenyo, D.; Beavis, R. C. *Proteomics* **2002**, *2*, 36−47.

10.1021/ac025747h CCC: $22.00 © 2002 American Chemical Society
Published on Web 09/12/2002

fragment ion masses between the observed spectrum and that predicted for the peptide but can reflect additional information, such as the difference in mass between spectrum parent ion and peptide. The SCOPE algorithm uses a probabilistic scoring scheme that can incroporate a priori knowledge regarding an experiment.[21]

A current challenge for high-throughput proteomics is to use database search results from large numbers of MS/MS spectra in order to derive a list of identified peptides and their corresponding proteins. This task necessarily entails distinguishing correct peptide assignments from false identifications among database search results. In the case of small datasets, this can be achieved by researchers with expertise manually verifying the peptide assignments to spectra made by database search programs. However, such a time-consuming approach is not feasible for high-throughput analysis of large datasets containing tens of thousands of spectra or when expertise is not available. Alternatively, researchers can attempt to separate the correct from incorrect peptide assignments by applying filtering criteria based upon database search scores and properties of the assigned peptides, such as the number of tryptic termini.[3−4,22] However, the numbers of rejected correct identifications and accepted false identifications that result from applying such filters are not known, nor how those numbers are affected by mass spectrometer, sample preparation, or spectrum quality. In addition, researchers often use different filtering criteria, making it particularly difficult to compare their results to one another.

In this work, we describe in detail a robust and accurate statistical model to assess the validity of peptide identifications made by MS/MS and database search. Each peptide assignment to a spectrum is evaluated with respect to all other assignments in the dataset, including necessarily some incorrect assignments. Employing database search scores and the number of tryptic termini of the assigned peptides, the method applies machine learning techniques to distinguish correctly from incorrectly assigned peptides in the dataset, and in so doing, computes for each peptide assignment to a spectrum a probability of being correct. We apply this method to SEQUEST database search results for ESI-MS/MS spectra generated from a control sample of 18 purified proteins.[23] Using this dataset with peptide assignments of known validity, we demonstrate that the computed probabilities are accurate and have high power to discriminate between correctly and incorrectly assigned peptides.

This statistical analysis promises to be of great value to high-throughput proteomics. Accurate probabilities with high discriminating power obviate the need for laborious manual verification of MS/MS database search results in the case of all but the most uncertain peptide identifications and enable filtering of data with predictable false identification error rates. This should facilitate the benchmarking of various mass spectrometer settings and experimental procedures to identify those that maximize the number of identifications per sample or per unit time. It can also serve as a common standard by which the results of different research groups, using different mass spectrometers, and even different database search software, can be compared. It is interesting to note that similar advantages have been realized in the field of large-scale DNA sequencing with the development of mathematical models for the estimation of errors in "raw" DNA sequence data.[24−26]

EXPERIMENTAL DATASETS

Tandem mass spectra used in this study were generated from 22 LC/MS/MS runs on a control sample composed of 18 purified proteins at a variety of concentrations, as previously described.[23] For each LC/MS/MS run, a control sample proteolyzed with trypsin was subjected to ESI-MS/MS. A training dataset of peptide assignments with known validity was obtained by searching these spectra with the SEQUEST analysis program[18] using a *Drosophila* peptide database[27] appended with sequences of the 18 control proteins. The database also included sequences of several human proteins, such as keratin, that are common sample contaminants. Since the low-resolution ESI ion trap mass spectrometer used to generate the MS/MS spectra cannot distinguish between $[M + 2H]^{2+}$ and $[M + 3H]^{3+}$ precursor ions, each spectrum was searched by SEQUEST against the database and assigned a peptide separately for each precursor ion charge. This analysis produced a training dataset with 18 496 peptide assignments to spectra of $[M + 2H]^{2+}$ ions and 18 044 peptide assignments to spectra of $[M + 3H]^{3+}$ ions. The 504 spectra of $[M + H]^{+}$ ions were omitted from the analysis. SEQUEST peptide assignments corresponding to the 18 control sample proteins or common contaminants could occur by chance and, hence, were manually scrutinized to determine whether they were correct. All peptide assignments corresponding to proteins other than the 18 in the control samples and the common contaminants were inferred to be incorrect. In total, 1687 peptide assignments to spectra of $[M + 2H]^{2+}$ ions and 1011 to spectra of $[M + 3H]^{3+}$ ions were determined to be correct.

A distinct test dataset of peptide assignments to spectra was obtained by searching the same spectra from above with SEQUEST using a human peptide database[27] appended with sequences of the 18 proteins of the control sample. This database is 2.5 times larger than that used for the training data, and it is expected to generate different SEQUEST score distributions and perhaps different correct and incorrect peptide assignments. Once again, all correct peptide assignments corresponding to one of the 18 control sample proteins or common contaminants were manually confirmed. Altogether, 1658 peptide assignments to spectra of $[M + 2H]^{2+}$ ions and 1001 to spectra of $[M + 3H]^{3+}$ ions were determined to be correct.

RESULTS AND DISCUSSION

Discriminant Function Analysis to Combine Database Search Scores. A useful statistical model to assess the validity of peptide assignments should enable discrimination between correct and incorrect peptide assignments on the basis of readily available information regarding the spectrum and assigned peptide. All database searching tools include with each peptide

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

[22] Han, D. K.; Eng, J.; Zhou, H.; Aebersold, R. *Nat. Biotechnol.* **2001**, *19*, 946−951.

[23] Keller, A.; Purvine, S.; Nesvizhskii, A. I.; Stolyar, S. et al. *OMICS* **2002**, *6* (2), 207−212.

[24] Lawrence, C. B.; Solovyev, V. V. *Nucleic Acids Res.* **1994**, *22*, 1272−1280.

[25] Ewing, B.; Hillier, L.; Wendl, M. C.; Green, P. *Genome Res.* **1998**, *8*, 175−185.

[26] Ewing, B.; Green, P. *Genome Res.* **1998**, *8*, 186−194.

[27] Extracted from ftp://ftp.ncicrf.gov/pub/nonredun/protein.nrdb.Z.

assignment to a spectrum a score or set of scores that reflects the likelihood that the given assignment is correct. Using Bayes' Law and denoting correct and incorrect peptide assignments as "+" and "−", respectively, the probability that a particular peptide assignment with database search scores, $x_1, x_2, ..., x_S$, is correct, $p(+|x_1, x_2, ..., x_S)$, can be computed as

$$p(+|x_1, x_2, ..., x_S) = \frac{p(x_1, x_2, ..., x_S|+)p(+)}{p(x_1, x_2, ..., x_S|+)p(+) + p(x_1, x_2, ..., x_S|-)p(-)} \quad (1)$$

where $p(x_1, x_2, ..., x_S|+)$ and $p(x_1, x_2, ..., x_S|-)$ are the probabilities of scores $x_1, x_2, ..., x_S$ among correctly and incorrectly assigned peptides, respectively, and $p(+)$ and $p(-)$ are prior probabilities of a correct and incorrect peptide assignment, respectively. The prior probabilities are the overall proportion of correct or incorrect peptide assignments in the dataset. To compute probabilities using eq 1, joint probability distributions for database search scores among correct and incorrect peptide assignments must be derived from training data with peptide assignments of known validity or learned from the data itself. In either case, this becomes more difficult as the number of database search scores, S, increases.

Discriminant function analysis can be used to combine together any number of database search scores, $x_1, x_2, ..., x_S$, into a single discriminant score that best separates training data into two groups by class, correct and incorrect peptide assignments.[28] The discriminant score, F, is a weighted combination of the database search scores, computed according to the discriminant function

$$F(x_1, x_2, ..., x_S) = c_0 + \sum_{i=1}^{S} c_i x_i \quad (2)$$

with constant c_0 and weights c_i derived in such a way that the ratio of between-class variation to within-class variation is maximized under the assumption of multivariate normality. Deriving the discriminant function requires training data with peptide assignments of known validity. The resulting discriminant score can be substituted into eq 1 in place of the original database search scores to enable tractable calculation of probabilities that retain as much discriminating power as possible using a single weighted combination of the scores,

$$p(+|F) = \frac{p(F|+)p(+)}{p(F|+)p(+) + p(F|-)p(-)} \quad (3)$$

where $p(+|F)$ is the probability that the peptide assignment with discriminant score F is correct, and $p(F|+)$ and $p(F|-)$ are the probabilities of F according to the discriminant score distributions among correct and incorrect peptide assignments, respectively.

Discriminant function analysis was applied to SEQUEST database search scores using the training dataset of search results with known validity for spectra generated from a control sample of 18 proteins. Spectra of poor quality, as determined by a score reflecting several spectrum properties,[29] were omitted during the derivation of the discriminant functions. Such "data cleaning" excluded 7070 spectra of $[M + 2H]^{2+}$ ions and 6995 spectra of $[M + 3H]^{3+}$ ions, >99% of which were incorrectly assigned, and led to functions with improved discrimination between correct and incorrect peptide assignments to spectra, even when applied to the entire "uncleaned" training dataset. Four SEQUEST scores were found to contribute significantly to discrimination and were included in the final discriminant function analysis: 1. cross-correlation (Xcorr), a measure based on the number of peaks of common mass between observed and expected spectra, and used as a primary criterion for peptide assignments; 2. ΔC_n, the relative difference between the first and second highest Xcorr score for all peptides queried from the database; 3. SpRank, a measure of how well the assigned peptide scored relative to those of similar mass in the database, using a preliminary correlation metric; 4. d_M, the absolute value of the difference in mass between the precursor ion of the spectrum and the assigned peptide. Separate analyses were applied to spectra of $[M + 2H]^{2+}$ and $[M + 3H]^{3+}$ precursor ions, since they give rise to different Xcorr and ΔC_n score distributions. Analysis of spectra of $[M + H]^+$ precursor ions was not pursued, given their poor representation in the training dataset.

Transformations of some input scores significantly improved the discrimination power of this approach. For example, increased discrimination between correct and incorrect peptide assignments was achieved by taking the natural log of Xcorr and SpRank, which reduces the variance of those scores, prior to discriminant analysis. A strong dependence of Xcorr on the length of assigned peptides was observed (see Supporting Information). Xcorr reflects the number of matches between ion fragments in the observed and predicted spectra and thus tends to be larger for assignments of long peptides, which have many fragment ions, than for assignments of short peptides. As a result, assignments of short peptides can be difficult to classify, since even correct assignments result in relatively small Xcorr scores. Length dependence of Xcorr scores was largely reduced by the transformation to Xcorr′,

$$\text{Xcorr}' = \begin{cases} \dfrac{\ln(\text{Xcorr})}{\ln(N_L)}, & \text{if } L < L_C \\ \dfrac{\ln(\text{Xcorr})}{\ln(N_C)}, & \text{if } L \geq L_C \end{cases} \quad (4)$$

where L is the length (number of amino acids) of the assigned peptide, N_L is the number of fragment ions expected for a peptide of length L, and L_C is a specified length threshold beyond which Xcorr is independent of peptide length, corresponding to number of fragment ions N_C. Values of N_L and L_C likely depend on the detectable mass range of the mass spectrometer. Nevertheless, N_L can be sufficiently approximated as $2L$ and $4L$ for the case of $[M + 2H]^{2+}$ and $[M + 3H]^{3+}$ precursor ions, respectively, since the major contributions to Xcorr are due to b and y ions. L_C was set at the peptide length beyond which Xcorr was found to be largely length-independent (15 and 25 for the case of $[M + 2H]^{2+}$ and $[M + 3H]^{3+}$ precursor ions, respectively). The transformation of Xcorr to Xcorr′ improved discrimination between correct and incorrect peptide assignments. It is interesting to note that a

[28] Tabachnick, B. G.; Fidell, L. S. *Using Multivariate Statistics*, 4th ed.; Allyn and Bacon: Needham Heights, MA, 2001.

[29] Nesvizhskii, A. I. Manuscript in preparation.

Table 1. Discriminant Functions Derived from Training Dataset Spectra of $[M + 2H]^{2+}$ and $[M + 3H]^{3+}$ Precursor Ions[a]

	$[M + 2H]^{2+}$		$[M + 3H]^{3+}$	
variable	coefficient	correlation	coefficient	correlation
Xcorr′	8.362	0.798	9.933	0.698
ΔC_n	7.386	0.746	11.149	0.806
ln SpRank	−0.194	−0.510	−0.201	−0.491
d_M	−0.314	−0.306	−0.277	−0.251
constant	−0.959		−1.460	

[a] The coefficients weighting each variable are indicated, as well as the correlation between each variable and the discriminant function given by the loading matrix, indicative of the contribution of each variable to discrimination. Contributions can range from none (correlation of 0) to complete (correlation of ±1).

similar transformation of FASTA and Smith−Waterman sequence similarity scores has been successfully employed to reduce their sequence-length dependence.[30]

The discriminant functions derived from the training data for spectra of $[M + 2H]^{2+}$ and $[M + 3H]^{3+}$ ions are listed in Table 1. From the magnitude of correlations between variables and discriminant score given by the loading matrix, it is evident that the SEQUEST Xcorr and ΔC_n scores contribute to most of the discrimination achieved between correctly and incorrectly assigned peptides. With the two scores, excellent linear separation between classes and normality was observed, thus validating the use of linear discriminant analysis for distinguishing correct and incorrect peptide assignments (see Supporting Information). Results with the training dataset support the effectiveness of the discriminant score. For example, among spectra of $[M + 2H]^{2+}$ precursor ion charge, 84% of correct peptide assignments had discriminant scores of 1.7 or greater, whereas 99% of incorrect peptide assignments had scores below that threshold.

Calculation of probabilities that peptide assignments are correct using eq 3 requires accurate models of discriminant score distributions. The observed discriminant score positive (correct peptide assignments) and negative (incorrect peptide assignments) distributions for spectra of $[M + 2H]^{2+}$ and $[M + 3H]^{3+}$ precursor ions in the training dataset are shown in Figure 1A and B, respectively. These distributions were obtained by placing spectra in bins of width 0.2 according to discriminant score, and counting the resulting total number in each bin. Several parametrized distributions were assessed for their fit to these distributions. A Gaussian distribution offers a close approximation to the observed discriminant score distributions among correct peptide assignments (Figure 1). Hence, the probability that a correct peptide assignment has discriminant score F can be computed according to a Gaussian distribution with calculated mean μ and standard deviation σ.

$$p(F|+) = \frac{1}{\sqrt{2\pi}\sigma}e^{-(F-\mu)^2/2\sigma^2} \qquad (5)$$

In contrast, the discriminant score negative distributions were noticeably asymmetric, having an extended right tail, and were satisfactorily modeled by a gamma distribution (Figure 1). The probability of a discriminant score F for an incorrect peptide assignment can thus be computed as

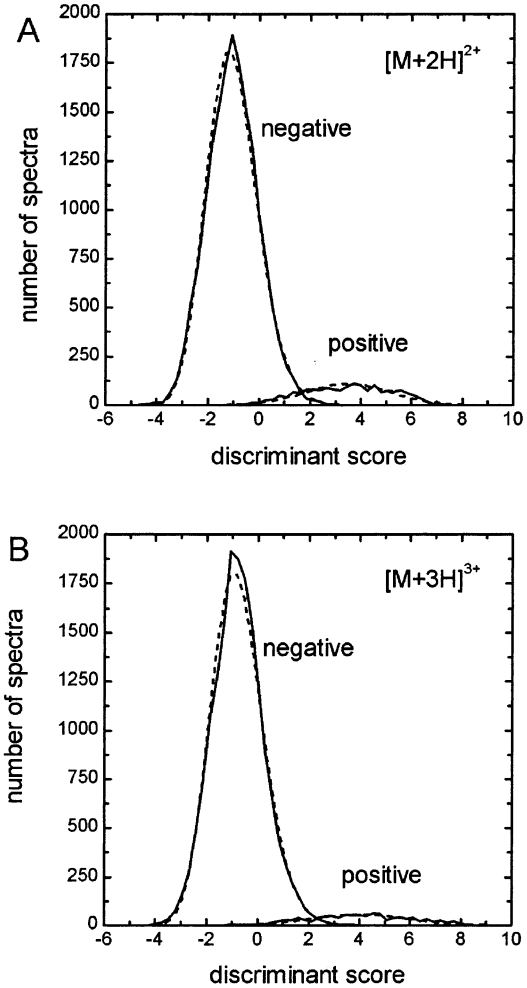

Figure 1. Training data discriminant score distributions. Discriminant score positive (correct peptide assignments) and negative (incorrect peptide assignments) disbributions for spectra of (A) $[M + 2H]^{2+}$ and (B) $[M + 3H]^{3+}$ precursor ions in the training dataset (solid line). Also shown are Gaussian and gamma distributions for the correct and incorrect peptide assignments, respectively, computed by the method of moments (dashed line).

$$p(F|-) = \frac{(F-\gamma)^{\alpha-1}e^{-(F-\gamma)/\beta}}{\beta^{\alpha}\Gamma(\alpha)} \qquad (6)$$

with parameter γ set below the minimum F value in the dataset, and parameters α and β computed using the method of moments.[31] Substituting the expressions for $p(F|+)$ and $p(F|-)$, along with the computed prior probabilities, into eq 3 should enable the calculation of accurate probabilities that peptides assigned to spectra in the training dataset are correct.

Computing Probabilities with NTT Distributions. When the enzyme trypsin is used to break proteins into small fragments amenable to MS/MS, database search of spectra can be performed in a constrained manner against only those peptides predicted to

[30] Pearson, W. R. *Protein Sci.* **1995**, *4*, 1145−1160.

[31] Johnson, N. L.; Kotz, S. *Continuous Univariate Distributions*. John Wiley & Sons: New York, 1994.

be products of trypsin digestion. This has the advantage of saving computation time by reducing the search space. Alternatively, database search can be performed in an unconstrained manner against all peptides in the database. This has the advantage of potentially identifying peptides that result from nontryptic cleavage, for example, by protease contaminants in the sample or from nonenzymatic cleavage during ionization for ESI-MS/MS.

In the case of unconstrained database searches, such as the SEQUEST searches used in this study, the number of tryptic termini (NTT) of peptides assigned to spectra is valuable information for assessing whether the assignments are correct and has often been used as partial criteria for accepting peptide assignments.[3–4,22] This number, which can be 0, 1, or 2, measures how many of the peptide termini, on the basis of amino acid sequence, are consistent with cleavage by trypsin. The NTT distributions (the relative numbers of peptides with 0, 1, or 2 tryptic termini) among correct and incorrect peptide assignments are sufficiently distinct to be of use for computing the probability that a peptide assignment is correct. Whereas the NTT distributions for incorrect peptide assignments are expected to be predominantly NTT = 0, reflecting the frequencies of amino acids recognized by trypsin (lysine and arginine) in the database used for search, those for correct peptide assignments are expected to be predominantly NTT = 2. In the training dataset, for example, of the incorrectly assigned peptides, 80% had NTT = 0, 19% had NTT = 1, and 1% had NTT = 2, whereas of the correctly assigned peptides, 3% had NTT = 0, 23% had NTT = 1, and 74% had NTT = 2. Taking into account NTT information in addition to the discriminant score when computing probabilities should result in improved discrimination between correct and incorrect results of unconstrained database searches. That is because even if two peptide assignments, one with NTT = 0 and the other with NTT = 2, have the same discriminant score, the latter is more likely to be correct since peptides with NTT = 2 are highly enriched among correct assignments.

Combining together information regarding NTT as well as discriminant score for a spectrum, the probability that a peptide assignment is correct, $p(+|F, \mathrm{NTT})$, can be expressed using Bayes' Law,

$$p(+|F, \mathrm{NTT}) = \frac{p(F, \mathrm{NTT}|+)p(+)}{p(F, \mathrm{NTT}|+)p(+) + p(F, \mathrm{NTT}|-)p(-)} \quad (7)$$

where $p(F, \mathrm{NTT}|+)$ and $p(F, \mathrm{NTT}|-)$ are the probabilities of F and NTT according to the positive and negative joint probability distributions, respectively. One expects the discriminant score and NTT distributions to be independent of one another, both for correct and incorrect peptide assignments, as long as the database search scores are not strongly dependent on the number of tryptic termini of peptides. This is true in the case of SEQUEST. Normalized discriminant score distributions are plotted in Figure 2 for correct and incorrect peptide assignments to spectra of $[\mathrm{M} + 2\mathrm{H}]^{2+}$ precursor ions in the training dataset, showing each separately for cases in which the assigned peptide has 0, 1, or 2 tryptic termini. It is evident that the discriminant score distributions are very similar for all values of NTT, verifying that the discriminant score and NTT distributions can be considered independent among correct and incorrect peptide assignments. Incorporating this independence assumption into eq 7 yields the simplified expression for computing the probability that a peptide assignment with discriminant score F and number of tryptic termini NTT is correct,

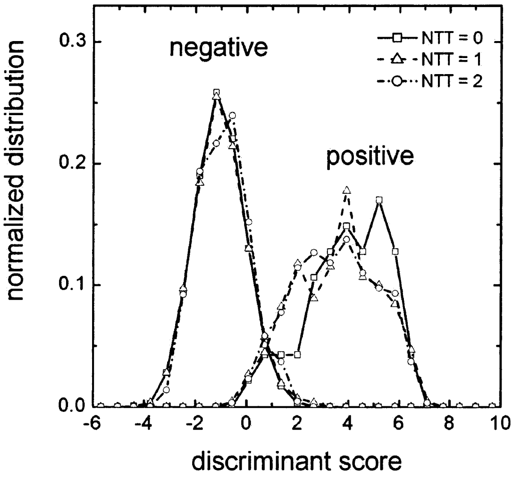

Figure 2. Independence of distributions of discriminant score and number of tryptic termini (NTT) among correct and incorrect peptide assignments. Normalized discriminant score positive (correct peptide assignments) and negative (incorrect peptide assignments) distributions for training data spectra of $[\mathrm{M} + 2\mathrm{H}]^{2+}$ precursor ions are shown separately for NTT values of 0, 1, and 2.

$$p(+|F,\mathrm{NTT}) = \frac{p(F|+)p(\mathrm{NTT}|+)p(+)}{p(F|+)p(\mathrm{NTT}|+)p(+) + p(F|-)p(\mathrm{NTT}|-)p(-)} \quad (8)$$

where $p(\mathrm{NTT}|+)$ and $p(\mathrm{NTT}|-)$ are the probabilities of NTT according to the distributions among correct and incorrect peptide assignments, respectively. If the database search is constrained to consider only fully tryptic peptides, then all assigned peptides will have NTT = 2, effectively reducing eq 8 to eq 3. The strategy outlined in this section for incorporating NTT information into computed probabilities should be generalizable to the use of enzymes other than trypsin, as well.

Mixture Model Distributions of Correct and Incorrect Peptide Assignments. Although the training data prior probabilities and discriminant score and NTT distributions of correct and incorrect peptide assignments could be used as a "global model" to compute probabilities that any test spectrum assignment is correct according to eq 8, such a model would not likely be robust enough to produce accurate probabilities for a wide variety of datasets. That is in part because the discriminant score distributions can vary significantly from dataset to dataset. For example, in the case of SEQUEST, Xcorr, one of the primary contributors to the discriminant score, is strongly affected by the levels of signal to noise in the spectra. Less than accurate computed probabilities would also result from variations in the NTT distributions of correctly assigned peptides, which are sensitive to the efficiency of sample trypsinization and the presence of protease contaminants in the sample. In addition, the NTT distributions of incorrectly assigned peptides, though more predictable, can nonetheless vary depending on the frequency of the amino acids recognized by trypsin (e.g. lysine and arginine) in the particular database used for search. Finally, a global model will not adequately reflect the expected variations in prior probabilities (the relative numbers of correctly and incorrectly assigned spectra) from dataset to dataset because of several factors, such as sample purity and spectral quality.

A robust alternative to using a global model based upon distributions derived from the training data is to derive the prior probabilities and discriminant score and NTT distributions among correct and incorrect peptide assignments empirically from each dataset as a mixture model using the expectation maximization (EM) algorithm.[32] In the mixture model, each spectrum contributes to distributions of both correct and incorrect peptide assignments in proportion to its computed probability of being correctly and incorrectly assigned, respectively. It is the task of the EM algorithm to learn the likelihood that each peptide assignment in the dataset is correct vs incorrect. At the outset, the EM algorithm makes initial estimates of the prior probabilities and discriminant score and NTT distributions among correct and incorrect peptide assignments. It then finds distributions that best fit the observed data by a two-step iterative process. In the first step, probabilities that spectra are correctly assigned, $p(+|F,$ NTT*)*, are calculated according to eq 8 using current estimates of the distributions. In the second step, the prior probabilities and discriminant score and NTT distributions among correct peptide assignments are computed using all spectra in the dataset, each weighted by the current estimate that it is correctly assigned, $p(+|F,$ NTT*)*. Similarly, the distributions among incorrect peptide assignments are computed, also using all spectra in the dataset, but in this case, each is weighted by the current estimate that it is incorrectly assigned, $(1 - p(+|F, \mathrm{NTT}))$.

For example, during the second step of the EM algorithm, the prior probability of a correct peptide assignment in a dataset of N spectra is computed as

$$p(+) = \frac{1}{N}\sum_{i=1}^{N} p(+|F_i, \mathrm{NTT}_i) \tag{9}$$

where F_i and NTT_i are the discriminant score and NTT value, respectively, for the peptide assignment to spectrum i. The discriminant score positive distribution is modeled as a Gaussian distribution, and the probability of a particular discriminant score, F, among correct peptide assignments is therefore given by eq 5 with mean μ and standard deviation σ, computed from all spectra as

$$\mu = \frac{1}{Np(+)}\sum_i p(+|F_i, \mathrm{NTT}_i)F_i \tag{10}$$

$$\sigma = \sqrt{\frac{1}{Np(+)}\sum_i p(+|F_i, \mathrm{NTT}_i)(F_i - \mu)^2} \tag{11}$$

Finally, the probability that a correct peptide assignment has a particular number of tryptic termini, $p(\mathrm{NTT}|+)$, is computed as the proportion of $p(+)$ contributed by all spectra j assigned peptides with that value of NTT.

$$p(\mathrm{NTT}|+) = \frac{1}{Np(+)}\sum_{\{j|\mathrm{NTT}_j=\mathrm{NTT}\}} p(+|F_j, \mathrm{NTT_j}) \tag{12}$$

The discriminant score and NTT distributions among incorrect peptide assignments are computed in an analogous manner, modeling the discriminant score negative distribution as a gamma distribution according to eq 6.

The EM algorithm two-step process of computing $p(+|F, \mathrm{NTT})$ for spectra using current estimates of the prior probabilities and discriminant score and NTT distributions and computing the prior probabilities and discriminant score and NTT distributions using current estimates of $p(+|F, \mathrm{NTT})$ is repeated until no significant changes in the distributions result. With each iteration, the mixture model distributions more closely match the observed data (see Supporting Information for an illustration of the EM algorithm applied to the training data). Upon termination of the algorithm, final probabilities that peptides assigned to spectra are correct are computed according to eq 8 using the learned distributions.

The EM algorithm is not guaranteed to find the most likely distributions of correct and incorrect peptide assignments in a dataset, so several constraints are imposed in order to guide the derivation of accurate distributions. For example, the discriminant score distribution for correct peptide assignments is initialized with that computed from the training data, which is expected to be similar, although not identical, to that of most datasets analyzed. The discriminant score distribution for incorrect peptide assignments in the mixture model is initialized with the distribution derived from data with $\mathrm{NTT} = 0$, since the majority of such cases are expected to be incorrect peptide assignments. The γ parameter value of the distribution is fixed during all subsequent rounds of the EM algorithm. In addition, the mixture model probabilities of all spectra with discriminant scores smaller than the initial discriminant score negative distribution mean are fixed at 0 to enforce that they be considered incorrectly assigned.

Combined Probabilities for $[\mathrm{M} + 2\mathrm{H}]^{2+}$ and $[\mathrm{M} + 3\mathrm{H}]^{3+}$ Precursor Ions. In the analysis thus far described, independent models are employed to estimate the probability that a peptide is assigned correctly to a spectrum of either an $[\mathrm{M} + 2\mathrm{H}]^{2+}$ or $[\mathrm{M} + 3\mathrm{H}]^{3+}$ precursor ion. This is justified when the precursor ion charge is known. In the case of high-resolution mass spectrometry, such as quadropole time-of-flight, the charge of the precursor ion can be determined by deconvoluting the isotopic pattern. In the case of low-resolution mass spectrometry, the precursor ion charge is generally not known, but it can in some cases be determined by computational methods.[33,34] In addition, database search applications can query each spectrum against a database separately for both $[\mathrm{M} + 2\mathrm{H}]^{2+}$ and $[\mathrm{M} + 3\mathrm{H}]^{3+}$ precursor ion cases and report only one peptide assignment per spectrum corresponding to the (putatively correct) charge state yielding the higher final score.

[32] Dempster, A.; Laird, N.; Revow, M. *J. R. Stat. Soc.* **1977**, *B39* (1), 1−38.

[33] Sadygov, R. G.; Eng, J.; Durr, E.; Saraf, A.; et al. *J. Proteome Res.* **2002**, *1* (3), 211−215.

[34] Perez, R. E.; Asara, J. M.; Lane, W. L. *Proceedings of the 50th ASMS Conference on Mass Spectrometry and Allied Topics*, Orlando, FL, 2002; in press.

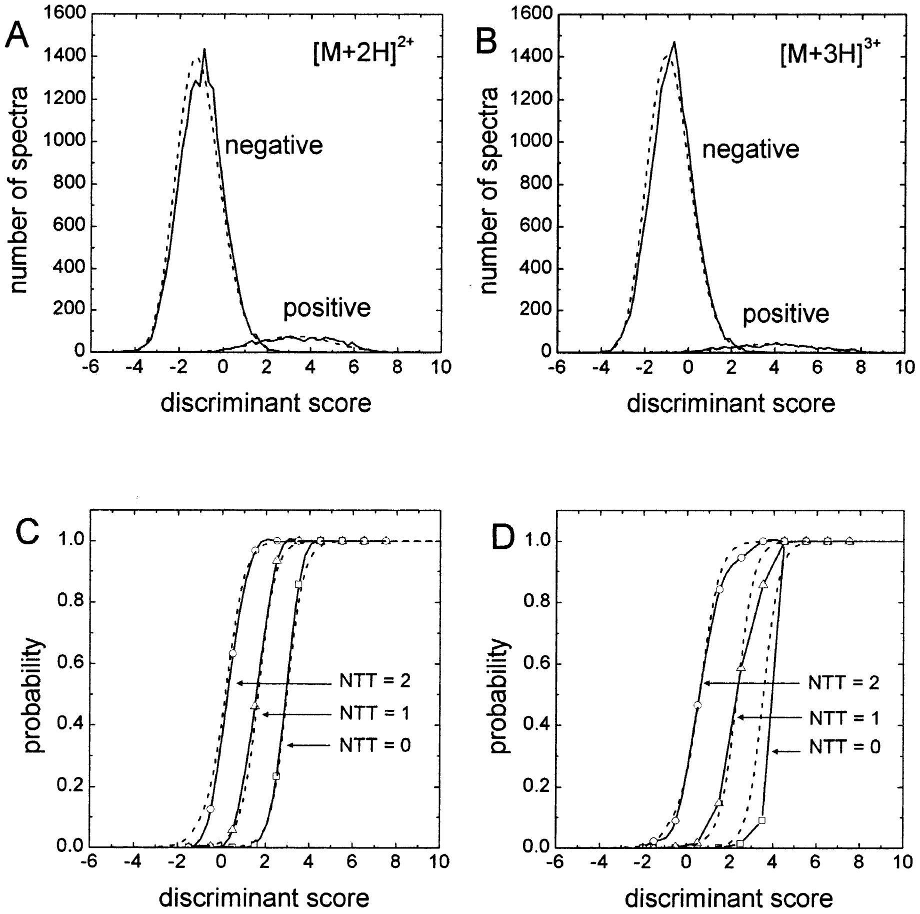

Figure 3. Test data discriminant score distributions and probabilities. Actual (solid line) and mixture model derived (dashed line) discriminant score positive (correct peptide assignments) and negative (incorrect peptide assignments) distributions for test data spectra of (A) [M + 2H]$^{2+}$ and (B) [M + 3H]$^{3+}$ precursor ions. Actual probability (fraction of database search results that are correct) (solid line) and computed probability (dashed line) plotted separately for peptide assignments with the number of tryptic termini (NTT) equal to 0, 1, or 2 for spectra of (C) [M + 2H]$^{2+}$ and (D) [M + 3H]$^{3+}$ precursor ions.

When the precursor ion charge of a spectrum is not known and database search results for both the [M + 2H]$^{2+}$ and [M + 3H]$^{3+}$ ion cases are reported, there must be a suitable means to reconcile the peptide assignments for each ion case. For example, the ion trap spectra of this study were searched by SEQUEST and assigned peptides for both [M + 2H]$^{2+}$ and [M + 3H]$^{3+}$ ion cases, yet each spectrum can realistically have a correct peptide assignment for at most one precursor ion case. Using the mixture models for [M + 2H]$^{2+}$ and [M + 3H]$^{3+}$ precursor ions described above, probabilities are computed that the peptide assignments are correct for each case independently, with no constraints that, for example, prevent probabilities of 1 being computed for peptide assignments to both precursor ion cases of the same spectrum.

As a means of combining together the results of both [M + 2H]$^{2+}$ and [M + 3H]$^{3+}$ precursor ion models in a realistic way in cases in which the ion charge is not known, probabilities computed by each model to the same spectrum are adjusted downward by the multiplicative factor, $(p^{2+} + p^{3+} - p^{2+}p^{3+})/(p^{2+} + p^{3+})$, where p^{2+} and p^{3+} are the independently computed probabilities that peptide assignments to a spectrum are correct for the [M + 2H]$^{2+}$ and [M + 3H]$^{3+}$ ion cases, respectively. This adjustment to the probabilities should have minimal effect on the great majority of spectra for which $p^{2+}p^{3+}$ is small. It effectively removes the unrealistic joint probability (assuming independence) that peptides assigned to a spectrum are correct for both [M + 2H]$^{2+}$ and [M + 3H]$^{3+}$ ion cases simultaneously. For example, in the case in which both p^{2+} and p^{3+} are initially 1 for a particular spectrum, they are each reduced to 0.5. This adjustment ensures in general that probabilities that peptide assignments are correct for both precursor ion cases of the same spectrum do not sum to more than unity.

Test Data Mixture Model Probabilities. The mixture model EM method was applied to a test dataset of combined MS/MS spectra generated from 22 different LC/MS/MS runs on a control sample of 18 proteins. These spectra were searched using SEQUEST against the human peptide database appended with sequences of the control proteins. This test dataset, similar to the training dataset described above, has SEQUEST peptide assignments with known validity and can be used to evaluate the mixture model method for computing probabilities that spectra are assigned correctly. Convergence of the EM algorithm was achieved within 10 iterations. Figure 3A,B shows the discriminant score positive and negative distributions derived by the mixture model EM method for spectra of [M + 2H]$^{2+}$ and [M + 3H]$^{3+}$ precursor ions, respectively. It is evident that they closely match the actual positive and negative distributions of the test dataset. Close correlation between actual and model-derived prior probabilities and NTT distributions was also observed.

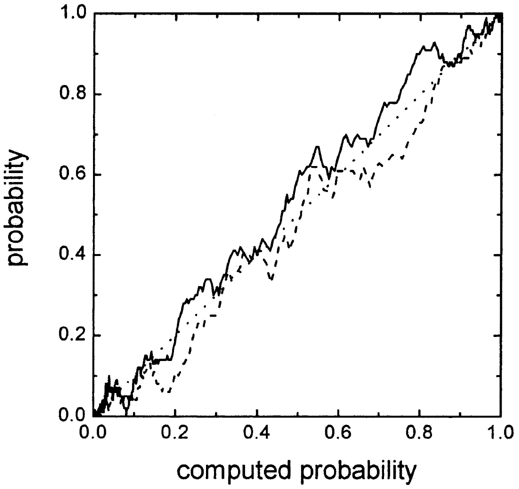

Figure 4. Accuracy of computed probabilities. The actual probability (fraction of peptide assignments that are correct) among spectra with indicated computed probabilities, determined either from a single mixture model using all 22 test data LC/MS/MS runs (solid line), or from 22 mixture models, each derived from an individual LC/MS/MS run (dashed line). The expected probability is also shown (dotted 45° line).

Probabilities that peptide assignments are correct were computed according to eq 8 using the positive and negative distributions estimated by the mixture model. Figure 3C,D shows for spectra of $[M + 2H]^{2+}$ and $[M + 3H]^{3+}$ ions, respectively, the computed probability and actual probability (fraction of database search results that are correct) plotted as a function of discriminant score separately for peptide assignments with NTT equal to 0, 1, or 2. Good agreement is evident for all values of NTT, justifying the assumption used to compute probabilities that for correct and incorrect peptide assignments, the discriminant scores and NTT are independent. As expected, assignments of peptides with NTT = 2 had much higher probabilities for any discriminant score relative to assignments of peptides with NTT = 0 or 1, reflecting the much higher proportion of peptides with NTT = 2 among correct assignments than among incorrect assignments.

The accuracy of computed probabilities is further demonstrated by plotting the actual probability that peptide assignments are correct as a function of computed probability for the combined test data spectra of $[M + 2H]^{2+}$ and $[M + 3H]^{3+}$ ions (Figure 4). Spectra were sorted by computed probability, and then the mean computed probability and actual probability were determined within a sliding window of 100 spectra. Good correspondence between computed and actual probabilities over most of the entire 0−1 range is evident, indicating that the computed probabilities are an accurate reflection of the likelihood that peptides are correctly assigned to spectra. This validates the mixture model EM method, and suggests that computed probabilities that peptides are correctly assigned to MS/MS spectra could be used effectively in a statistical model to estimate the likelihood that proteins corresponding to those peptides are present in the sample.

To evaluate the mixture model EM analysis method on smaller datasets, it was applied to database search results for each of the 22 test data LC/MS/MS runs separately to derive 22 different mixture models. LC/MS/MS runs typically produce only 1000−2000 spectra. Computed probabilities from the individual models were then combined and sorted to enable calculation of a mean computed probability and actual probability within a sliding window of 100 spectra, as previously described. Figure 4 shows that probabilities computed for individual LC/MS/MS runs are nearly as accurate as those computed from a single model of the combined data. Interestingly, even more accurate computed probabilities were obtained by removing poor quality spectra from the datasets prior to analysis.[29] Such data cleaning resulted in discriminant score negative distributions that were more faithfully modeled by a gamma distribution. These results demonstrate that the mixture model is generally applicable to database search results for small numbers of spectra obtained from individual LC/MS/MS runs.

Sensitivity and Error Rates with Minimum Computed Probability Thresholds. A useful statistical model, in addition to yielding accurate probabilities, should enable good discrimination between correct and incorrect database search results. An ideal model would enable complete separation between correct and incorrect peptide assignments, the former assigned probabilities close to unity, the latter, close to 0. In practice, one can accept all peptide assignments having a computed probability greater than or equal to a user-specified minimum threshold. A relatively low minimum threshold can be used to ensure high sensitivity (fraction of all correct peptide assignments accepted), yet often with the cost of a higher false identification error rate (fraction of accepted peptide assignments that are incorrect). Alternatively, a relatively high minimum threshold can be used to achieve a lower error rate at the expense of decreased sensitivity. The optimal tradeoff between sensitivity and error will depend on the relative importance given to each by the user.

The sensitivity/error rate tradeoff for the combined test dataset is illustrated in Figure 5A. Each point along the curve represents the results of using a different minimum probability threshold to accept all peptide assignments with computed probabilities at least as great. The bottom right corner (denoted with an asterisk) corresponds to an ideal data filter with 100% sensitivity and 0% error. For comparison, the plot also shows the results of using several conventional means of filtering data based upon SEQUEST scores and NTT. It is evident that data filters based on the probabilities computed from the model outperform each conventional filter, achieving much higher sensitivity and, hence, a greater number of correct identifications, for the same rate of error. For example, employing a conservative set of SEQUEST score and NTT thresholds (no. 5 in Figure 5A: Xcorr ≥ 2, $\Delta C_n \geq 0.1$, SpRank ≤ 50, NTT = 2) yields 61% sensitivity with an error of 2.5%, whereas employing a minimum probability threshold ($p \geq 0.65$) yields 89% sensitivity with the same level of error. Filtering data based upon computed probabilities in this case therefore confers 46% more correct peptide identifications than filtering data based upon SEQUEST scores and NTT.

To take advantage of good discrimination between correct and incorrect peptide assignments, a user must choose an appropriate minimum probability threshold to employ for a particular dataset. Accurate probabilities generated by the mixture model enable the expected sensitivity, ⟨Sens⟩, and expected false identification error

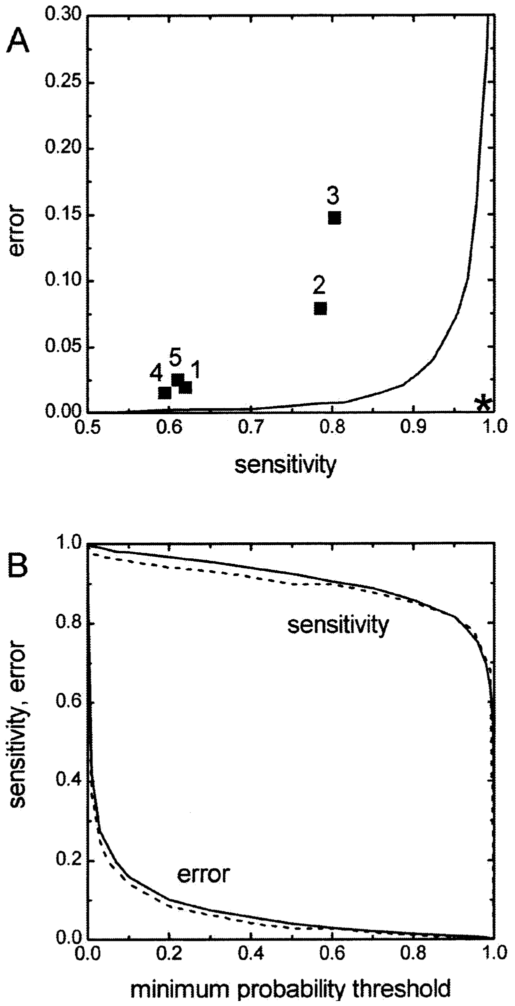

Figure 5. Sensitivity and false identification error rates using minimum computed probability thresholds. (A) Sensitivity/error rate tradeoff employing thresholds based upon computed mixture model probabilities (solid line). Also shown are results of using conventional filtering criteria based upon SEQUEST scores and the number of tryptic termini (NTT) of assigned peptides: (1) ref 4; (2) Xcorr $[M + 2H]^{2+}$ ions ≥ 2, Xcorr $[M + 3H]^{3+}$ ions ≥ 2.5, $\Delta C_n \geq 0.1$, SpRank ≤ 50, NTT ≥ 1; (3) Xcorr ≥ 2, $\Delta C_n \geq 0.1$, SpRank ≤ 50, NTT ≥ 1; (4) Xcorr $[M + 2H]^{2+}$ ions ≥ 2, Xcorr $[M + 3H]^{3+}$ ions ≥ 2.5, $\Delta C_n \geq 0.1$, SpRank ≤ 50, NTT = 2; (5) Xcorr ≥ 2, $\Delta C_n \geq 0.1$, SpRank ≤ 50, NTT = 2. The result of using an ideal data filter conferring 100% sensitivity and 0% error is indicated by an asterisk in the lower right corner. (B) Observed (solid line) and model-predicted (dashed line) sensitivity and error rate as a function of minimum computed probability threshold.

rate, ⟨Err⟩, for any minimum probability threshold value, p_T, to be calculated,

$$\langle \text{Sens} \rangle = \sum_{\{i|p_i \geq p_T\}} p_i \Big/ \sum_i p_i \tag{13}$$

$$\langle \text{Err} \rangle = \sum_{\{i|p_i \geq p_T\}} (1 - p_i) \Big/ \sum_{\{i|p_i \geq p_T\}} 1 \tag{14}$$

where p_i is the probablity that the peptide assigned to spectrum i is correct, and the indicated sums are either over all spectra i, or over only those spectra with $p_i \geq p_T$. The sensitivity and error rates predicted by the model for the test dataset agree well with those observed (Figure 5B) and can be used to select the minimum probability threshold that achieves the optimal tradeoff between the two or a specified false identification error rate. The use of such a threshold greatly facilitates high-throughput analysis of peptide identifications made by MS/MS and database search.

Future Work. The discriminant functions used throughout this work were derived using spectra produced on a single ESI ion trap mass spectrometer from a control sample of known protein components. It is possible that these functions may not be optimal for data produced from all other types of mass spectrometers (e.g., ESI-qTOF, MALDI-qTOF). One solution is to derive new discriminant functions specifically optimized for use with each different spectrometer type. This could be achieved using search results for spectra generated from a control sample of known components, as described in this work. Alternatively, it may be possible to learn an optimal discriminant function from conventional data by initially applying the mixture model EM method with the suboptimal discriminant score and then using the resulting estimated probabilities that peptide assignments are correct to weight the data in a derivation of a new discriminant function that optimally separates the data on the basis of those estimates. This procedure can be iterated until no significant change in the discriminant function results. We are currently exploring the feasibility of such an approach.

As described above, the mixture model discriminant score negative distribution is initialized using data with NTT = 0, which are predominantly incorrect. This is not possible for results of a database search with constraints on the minimum number of tryptic termini of assigned peptides. As an alternative means of guiding the derivation of the discriminant function negative distribution when data with NTT = 0 are not present, a set of "known incorrect peptide assignments" of equal number to the data, can be included in the mixture model during iterations of the EM algorithm. These incorrect assignments can be obtained, for example, by searching the dataset with SEQUEST using a database in which all protein sequences in the original database have been reversed. Such a database preserves the tryptic peptide length distribution of the original, yet ensures that all peptide assignments resulting from its use are incorrect (provided that all those present by chance in the original database are removed). Preliminary results using this strategy are encouraging.

The mixture model EM method is currently being extended to analyze peptide assignments to spectra of $[M + H]^+$ precursor ions. It can also be adapted to utilize additional information when available, such as the number of missed tryptic cleavages[35] and the expected pI of the peptide,[36] or in the case of experiments employing various chemical labels such as the ICAT reagent, the presence or absence of labeled amino acids in the assigned peptide or additional features of the spectrum related to that label. In addition, this approach is not specific to SEQUEST, but can in principle be applied to the results of any spectrum database search analysis.

Computed probabilities that peptides are correctly assigned to MS/MS spectra can be used to estimate the likelihood for the

[35] Parker, K. C. *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 22−39.

[36] Zuo, X.; Speicher, D. W. *Proteomics* **2002**, *2*, 58−68.

presence of proteins corresponding to those peptides in a sample.[37] Interestingly, correct peptide assignments, more than incorrect ones, tend to correspond to "multihit" proteins, those to which other correctly assigned peptides correspond. This trend is particularly pronounced for large datasets or samples of low complexity and can be exploited by adjusting the probabilities of assigned peptides to reflect whether their corresponding proteins are "multihit" in the dataset.

CONCLUSIONS

The statistical model described in this work enables high throughput analysis of MS/MS database search results and can serve as a useful standard by which the results of different research groups, using different mass spectrometers, and even different database search software, can be compared. It requires minimal user interaction and adds little execution time in addition to the database search. It eliminates the need to manually analyze database search results to assess whether they are correct. The probabilities computed by this analysis can instead be used effectively to identify correct peptide assignments and filter data with predictable false identification error rates. They also serve as useful inputs for estimating the likelihood of the presence of corresponding proteins in the sample.

Currently, the software implementing the mixture model EM analysis consists of a series of stand-alone Perl scripts that are run on Linux. It is available to the public as open source at http://www.systemsbiology.org/research/software.html. A Windows version of the software is also planned. Although the current version employs a discriminant function suitable for SEQUEST search results, its modular nature facilitates substituting alternative discriminant functions produced for scores of other database search tools, or simply single scores of such tools, to enable statistical analysis of a wide variety of database search applications.

ACKNOWLEDGMENT

This work was supported by DOE Grant DE-FG03-01ER63218 to E.K. and by NCI Grants R33CA93302 and R33CA84698, sponsored research agreement from Oxford GlycoSciences, and a gift from Merck & Co., Inc. to R.A.. We thank Andy Siegel, Jimmy Eng, and Jeff Ranish for helpful discussions. We also appreciate contributions from Dave Goodlett, Sergey Stolyar, and Sam Purvine regarding the datasets of spectra generated from the control sample and the comments of Jamie Sherman regarding the manuscript.

SUPPORTING INFORMATION AVAILABLE

Data illustrating the length independence of Xcorr′, the contributions of Xcorr′ and ΔC_n to the discriminant function, and the mixture model EM algorithm applied to the training dataset spectra of $[M + 2H]^{2+}$ precursor ions. This material is available free of charge via the Internet at http://pubs.acs.org/.

Received for review May 2, 2002. Accepted August 9, 2002.

AC025747H

[37] Nesvizhskii, A. I.; Keller, A.; Aebersold, R.; Kolker, E. Manuscript in preparation.