Anal. Chem. **2003,** *75,* 768−774

A Method for Assessing the Statistical Significance of Mass Spectrometry-Based Protein Identifications Using General Scoring Schemes

David Fenyö and Ronald C. Beavis*

Genomic Solutions (Canada) Inc., 460−70 Arthur Street, Winnipeg, Manitoa R3B 1G7, Canada

This paper investigates the use of survival functions and expectation values to evaluate the results of protein identification experiments. These functions are standard statistical measures that can be used to reduce various protein identification scoring schemes to a common, easily interpretably representation. The relative merits of scoring systems were explored using this approach, as well as the effects of altering primary identification parameters. We would advocate the widespread use of these simple statistical measures to simplify and standardize the reporting of the confidence of protein identification results, allowing the users of different identification algorithms to compare their results in a straightforward and statistically significant manner. A method is described for measuring these distributions using information that is being discarded by most protein identification search engines, resulting in accurate survival functions that are specific to any combination of scoring algorithms, sequence databases, and mass spectra.

Protein identification has become an important application of mass spectrometry to biological research.[1−3] These identifications are based on the comparison of an experimentally determined list of masses and a database of protein (or translated nucleotide) sequences. These masses can be a list of parent ion masses or a list of fragment ion masses derived from a parent ion by tandem mass spectrometry. This paper will address the results from tandem mass spectrometry experiments, but the arguments and findings may also be applied to protein identifications obtained by simple mass spectrometry.

Numerous algorithms are available that allow a user to enter a parent ion mass, a list of fragment ion masses, and search parameters that are used to find the peptide sequence in a sequence database that most closely matches those masses, given the search parameter set.[4] The matching is based on the known peptide bond fragmentation reactions.[5] These algorithms can be separated into two general classes: those that require interpretation of the mass list (class A) and those that require no interpretation (class B). Class A algorithms require the manual or automated determination of one or more stretches of unambiguous sequence by "de novo" sequencing methods.[6,7] The database search can then be performed with the knowledge that this sequence should be present in any matching peptide. Class B algorithms do not require the discovery of any sequence before a search is performed.[8,9] Class B algorithms compare all of the masses on the list with all peptide sequences. The parent ion mass is commonly used to limit the number of potential sequences considered. Both types of algorithms have advantages and disadvantages, but either type can be used to effectively identify protein sequences.

All of these algorithms calculate a "matching score" that is used as a measure of how closely a given peptide sequence matches the masses. Examples of these scores are correlation factors,[8] the number of ions that match a peptide sequence,[9] *k*-similarity statistics,[10] and calculated probability factors.[7] The score is dependent on the number and relative intensity of sequence-specific ions present in the mass list and the algorithm parameters used for the particular calculation.

Theoretical Interpretation of Peptide−Spectrum Matching Scores. In a typical protein identification experiment, a protein identification algorithm will be used to prepare a list of spectrum-to-peptide matching scores. The list is then presented to the investigator with the potential peptide sequences listed in order

* Corresponding author. E-mail: rbeavis@proteide.net.

(1) Ho, Y.; Gruhler, A.; Heilbut, A.; Bader, G. D.; Moore, L.; Adams, S. L.; Millar, A.; Taylor, P.; Bennett, K.; Boutilier, K.; Yang, L.; Wolting, C.; Donaldson, I.; Schandorff, S.; Shewnarane, J.; Vo, M.; Taggart, J.; Goudreault, M.; Muskat, B.; Alfarano, C.; Dewar, D.; Lin, Z.; Michalickova, K.; Willems, A. R.; Sassi, H.; Nielsen, P. A.; Rasmussen, K. J.; Andersen, J. R.; Johansen, L. E.; Hansen, L. H.; Jespersen, H.; Podtelejnikov, A.; Nielsen, E.; Crawford, J.; Poulsen, V.; Sorensen, B. D.; Matthiesen, J.; Hendrickson, R. C.; Gleeson, F.; Pawson, T.; Moran, M. F.; Durocher, D.; Mann, M.; Hogue, C. W.; Figeys, D.; Tyers, M. *Nature* **2002**, *415*, 180−183.

(2) Gavin, A. C.; Bosche, M.; Krause, R.; Grandi, P.; Marzioch, M.; Bauer, A.; Schultz, J.; Rick, J. M.; Michon, A.M.; Cruciat, C. M.; Remor, M.; Hofert, C.; Schelder, M.; Brajenovic, M.; Ruffner, H.; Merino, A.; Klein, K.; Hudak, M.; Dickson, D.; Rudi, T.; Gnau, V.; Bauch, A.; Bastuck, S.; Huhse, B.; Leutwein, C.; Heurtier, M. A.; Copley, R. R.; Edelmann, A.; Querfurth, E.; Rybin, V.; Drewes, G.; Raida, M.; Bouwmeester, T.; Bork, P.; Seraphin, B.; Kuster, B.; Neubauer, G.; Superti-Furga, G. *Nature* **2002**, *415*, 141−147.

(3) Andersen, J. S.; Lyon, C. E., Fox, A. H.; Leung, A. K.; Lam, Y. W.; Steen, H.; Mann, M.; Lamond, A. I. *Curr. Biol.* **2002**, *12*, 1−11.

(4) Rappsilber, J.; Mann, M. *Trends Biochem. Sci.* **2002**, *27*, 74−78.

(5) Aebersold, R.; Goodlett, D. R. *Chem. Rev.* **2001**, *101*, 269−295.

(6) Shevchenko, A.; Sunyaev, S.; Loboda, A.; Shevchenko, A.; Bork, P.; Ens, W.; Standing, K. G. *Anal. Chem.* **2001**, *73*, 1917−1926.

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

(8) Yates, J. R., 3rd; Eng, J. K.; McCormack, A. L.; Schietz, D. *Anal. Chem.* **1995**, *67*, 1426−1436.

(9) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **199**, *20*, 3551−3567.

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

10.1021/ac0258709 CCC: $25.00
Published on Web 01/10/2003

of decreasing score. Comparing the scores produced by any of the commonly available software packages on any basis other than the relative rank of a particular peptide in the list of peptides is difficult. As a result, practitioners of protein identification have evolved sets of rules that they apply to identifications in addition to the peptide rank and score to determine the validity of a protein identification. For the purposes of this discussion, a peptide sequence that is the same as the sequence of the real peptide that generated a particular tandem mass spectrum will be referred to as the "valid" sequence for that spectrum. All other peptide sequence assignments to a mass spectrum will be referred to as "stochastic".

One commonly used method to assess the validity of an assignment is to carefully examine the mass spectrum and determine how many of the ions were matched with the sequence. The user then uses his intuition and experience to determine whether the pattern of sequence ions assigned corresponds to a "reasonable" identification. The mass spectrum is also inspected to determine the signal-to-noise ratio of the data: high scores can be produced by purely stochastic assignments if there is a large amount of noise in the spectrum. This type of evaluation cannot be automated, and it depends on the skill of an individual to correctly interpret a complex set of patterns with an understanding of the instrumentation used to produce the original data.

Another strategy for determining the quality of a match to a particular peptide is to compare the ranked lists obtained from different algorithms. The user then applies some set of heuristic rules to compare the ranks of the peptides in the resultant lists. The simplest example of this approach would be to assume that if a peptide is top ranked by two algorithms, it may be assumed to be a correct assignment. This strategy is based on the assumption that different algorithms produce matches via methods that are sufficiently different that the highest scoring peptides resulting from stochastic matches to the data will be ranked differently by the algorithms, while peptides that are valid matches to the data will be ranked similarly. There is no reason to believe that this assumption is true. The opposite is probably true: both stochastic and valid assignments are likely to be similarly ranked by any two algorithms in the same class. Any two algorithms within a given class are based on similar assumptions, even though the details of the scoring system may be different. The practitioner is forced to use sets of algorithms that give similar results for trial spectra, which indicates that the algorithms must be formally equivalent to some degree. This formal equivalence means that comparing peptide rankings between equivalent algorithms to determine the quality of an assignment will lead to frequent nonvalid stochastic assignments.

A difficulty in comparing the scores resulting from different algorithms is that they do not necessarily have any physical interpretation. A good scoring system does not necessarily have any theoretical foundation: it is judged as good if it scores valid sequences significantly different from all other sequences. Therefore, any practically useful comparison between scores should be a statistic that can be applied to a variety of scoring systems and has a simple interpretation that is relevant to protein identification experiments.

A statistic that has these properties and that has found wide acceptance in bioinformatics is the expectation value.[11−13] An expectation value can be defined for any valid scoring system that has the characteristic of the score being a maximum when a valid match is obtained. Let x represent a score for a mass spectrum **S**; then the survival function, $s(x)$, for a discrete stochastic score probability distribution, $p(x)$, can be defined:[14]

$$s_{j(x)} = \Pr(X > x) = \sum_{i=j(x)}^{\infty} p_i \tag{1}$$

where $\Pr(X > x)$ is the probability that the spectrum's score will have a value greater than x by random matching with sequences in a particular database, **D**, and $p_{j(x)}$ is the discrete probability that best corresponds to a score of x. The expectation value $e(x)$ for **S** on **D** is then defined as

$$e_{j(x)} = n s_{j(x)} \tag{2}$$

where n is the number of sequences scored. The expectation value has the following simple interpretation: given **S** and **D**, $e(x)$ is the number of peptides that would be expected to have scores of at least x. For example, if a score x_0 has an expectation value $e(x_0) = 10$, then one would have a score of at least that value 10 times for every replicate of the experiment. An expectation value $e(x_0) = 1 \times 10^{-6}$ implies that the experiment would have to be repeated 1 million times before a score that high would be obtained by chance. Alternately, $e(x_0) = 1 \times 10^{-6}$ could be interpreted as the requirement that the sequence database searched would have to contain 1 million times as many sequences to obtain a score that high by chance.

These statistics can be applied to any type of scoring system for all of the peptide sequences that fit the primary parameters of a search. A primary search parameter that is common to all current scoring systems is the absolute value of the deviation, d, between the observed parent ion mass (m) and the calculated mass of a peptide sequence (m_c)

$$d = |m - m_c| \tag{3}$$

All sequences that do not fall within the appropriate d range specified for a search are not considered ($x = -\infty$).[11] This type of scoring is used in all tandem mass spectrometry identification schemes in common use currently, although there is usually no explicit score associated with sequences that do not have an appropriate intact mass. This scoring has the advantage of speeding up the calculation, because simply screening the parent ion mass removes a sequence from consideration without having to perform any further calculations.[7,8]

The only difficulty with applying eqs 1 and 2 to protein identification scores is that the stochastic probability distribution $p(x)$ is not readily available in a general, parametrized form. The distribution can be obtained by Monte Carlo methods, but the

(11) Karlin, S.; Altschul, S. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 2264−2268.
(12) Karlin, S.; Altschul, S. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 5873−5877.
(13) Mackey, A. J., Haystead, T. A. J.; Pearson, W. R. *Mol. Cell. Proteomics* **2002**, *1*, 139−147.
(14) Filliben, J. J.; Heckert, A. Exploratory Data Analysis. In *Engineering Statistics Handbook*; Croarkin, C., Tobias, P., Eds.; NIST: Washington DC; http://www.itl.nist.gov/div898/handbook/index.htm.

specific distribution applicable to a particular mass list and set of search parameters (including the sequence database) would be very time-consuming to calculate for each protein identification experiment. An alternative approach is to construct a frequency histogram of all peptide scores while a search is performed. During a search of a large sequence database, almost all scores will be the result of a random match: only one (if any) of the scores will be a valid match. The discrete valued frequency function $f(x)$ can be converted into $p(x)$ by simple normalization

$$p_i = f_i / N \tag{4}$$

where N is the number of peptide sequences used to create the histogram

$$N = \sum_{i=0}^{\infty} f_i \tag{5}$$

Applying eqs 1 and 5 makes it possible to estimate $p(x)$ and $s(x)$ for any scoring system with any combination of data, search parameters, and sequence. Once $s(x)$ is known, it can then be fit to an analytical function and its value extrapolated for any value of x. More importantly for the purposes of this paper, the survival functions for different scoring systems can be compared to evaluate their relative merit.

Experimental Determination of $p(x)$ and $s(x)$. All calculations were performed using modified versions of the database search engine Sonar.[15] Sonar uses vector alignment and spectrum dot products[16] as a method of calculating correlations between peptide sequences and mass spectra, which are represented as vectors with high dimensionality.[17] The calculation of the dot product is done in such a way as to also yield detailed assignments of sequence-specific ions at the same time as the sequence-spectrum correlation statistics are obtained.

The masses derived from calculating all of the potential peptide sequence-specific ions (a, b, y, and their corresponding −18 and − 17 neutral loss products) were represented in a vector **P** of n intensities P_i, where n is the mass of the parent ion divided by the accuracy of mass measurement. Possible ions were represented by $P_i = 1$, and all other values were 0. The spectrum to be compared was represented as an n point vector **I**, where the values of I_i corresponded to the intensity of an observed fragment ion (normalized to a maximum of 100) or 0 if no fragment ion was observed. The correlation score (indicated below as system I), x_I, was calculated as follows:

$$x_{\mathrm{I}} = \sum_{i=0}^{n} I_i P_i \tag{6}$$

Two alternative scoring schemes (systems II and III, respectively) were used for comparison: these schemes include extra terms that account for the number of assigned b or y ions (n_b and n_y, respectively),

$$x_{\mathrm{II}} = x_I(n_b! n_y!) \tag{7}$$

and

$$x_{\mathrm{III}} = x_I \exp(n_b + n_y) \tag{8}$$

While these additional factors may seem odd initially, consider the simple question: What is the form of $x(n_y)/x(n_y - 1)$? Consider the case of an 11-residue peptide sequence, **P**, and two spectra, **A** and **B**, such that spectrum **A** contains 10 peaks of equal intensity that can be assigned to y-ions predicted for **P**, while **B** contains only one peak that can be assigned to **P**. The correlation score assumes that the score should be a simple linear function of the number of assignable ions, implying that **A** is ∼10× better than **B** as proof of the existence of **P**. Intuitively, this relationship is too weak, as a spectrum with 10 assignable y-ions would be considered as good an identification as possible, while a single ion would be considered to have little value as proof. Assuming an underlying hypergeometric distribution for a valid match, $x(n_y)/x(n_y - 1) \approx n_y$, implying that **A** would be ∼(10!)× better than **B,** as proof of the existence of **P**. The exponential score would place the factor at ∼(e^{10})×. These larger factors are intuitively more reasonable and have the effect of moving valid scoring events to higher scores than the predominant stochastic distribution.

The SALSA score (system IV) was included as an additional scoring scheme: calculated based on the definition of SALSA scoring:[18]

$$x_{\mathrm{IV}} = K(\prod_{i=1}^{k} I_i)^{1/k} \tag{9}$$

where K is the number of detected ions that correspond to calculated peptide fragment ions and k is the total number of ions expected from the calculation. If an ion is not detected, the threshold of detection is used for T_i. As all signals are normalized to a maximum value of 100, the threshold value was set to 1 for all spectra considered here. The original description of the scoring scheme suggested that it was derived empirically from experimental data sets.

MATERIALS AND METHODS

All software was coded in C++, using Microsoft Visual Studio, version 6. The applications were run using the Common Gateway Interface on either a single processor Pentium III (600 MHz) or a single processor Pentium 4 (1.6 GHz) computer, through the Apache Web Server. All of the searches were obtained by searching fragment ion spectra against the National Center for Bioinformatics nonredundant protein sequence database (nr).

A single "real" tandem mass spectrum was selected to use as an exemplar for the calculation of the distributions below. This tandem spectrum was chosen from a reversed-phase high-performance liquid chromatography analysis of a protein tryptic digest performed on a Thermo-Finnigan LCQ ion trap mass

(15) Field, H. I.; Fenyö, D.; Beavis R. C. *Proteomics* **2002**, *2*, 36−47.
(16) Stark, H.; Woods, J. W. *Probability and Random Processes with Applications to Signal Processing*, 3rd ed.; Prentice Hall, Saddle River, NJ, 2002.
(17) Wan, K. X.; Vidavsky, I.; Gross, M. L. *J. Am. Soc. Mass Spectrom*. **2002**, *13*, 85−88.
(18) Liebler, D. C.; Hansen, B. T.; Davey, S. W.; Tiscareno, L.; Mason, D. E. *Anal. Chem*. **2002**, *74*, 203-210.

spectrometer. The spectrum yielded the same peptide sequence identification from the commercial spectrum search engines Mascot and Sonar and that sequence was assumed to be correct as eight other peptides from the protein (type I human keratin) were also identified. It should be stressed that the particulars of the mass spectrum are not important in the context of this paper: this single spectrum was chosen to be illustrative of the general approach. It was decided that using a single spectrum would emphasize the main theme of the paper, which is how to evaluate the results from a single spectrum when the scoring algorithms and search parameters were changed.

RESULTS AND DISCUSSION

The application of specialized sequence search engines to mass spectrometry has become a relatively well established technique for matching peptide and protein sequences to experimental results. The core analytical question that underlies much of this research may be summarized as follows: "Does this mass spectrum correspond unambiguously with a known protein or peptide sequence, yes or no?" If the answer is "yes", the researcher must then determine some type of confidence measure that the positive result is not simply a statistical coincidence, but that it is a valid match.

The range of scoring systems represented by eqs 6−9 underlines the difficulties for the average user in directly interpreting scores derived from tandem mass spectrometry protein identification search engines. These scoring systems have the same goal: to minimize the score obtained from stochastic matches while maximizing the score obtained from valid identifications. They approach the problem using different mathematical devices, however, making direct comparison of the scores difficult. It would be convenient to have a purely theoretical form for the underlying statistical distributions that would allow the directly calculation of the probability and expectation value for a particular score. Unfortunately, it is difficult to provide such a theoretical distribution in practice. There are so many possible parameters that would affect the distribution that it is not practical to attempt to model them all and still have an accurate representation of the experimental situation.

It is possible, however, to experimentally determine these distributions by accumulating a simple histogram during any peptide identification calculation. To illustrate this type of calculation, consider the case where a tandem mass spectrum has been generated using a peptide of known sequence. Using the formalism and algorithms described above, it is possible to directly measure the distribution of scores that result from peptides that do not correspond to the sample peptide that generated a particular tandem spectrum. This distribution will be referred to as the "stochastic" distribution, as it represents the scores possible from peptides with sequences different from the peptide that was actually analyzed, from a selected sequence database.

The sequence database used is important, as it does not contain a truly random distribution of peptide sequences. Calculations performed with truly random peptide sequences may yield results quite different from those measured from the relevant sequence database. The sequence database is assumed have the following properties: (a) it is large (meaning that it has many peptides that satisfy step 1 below); (b) if redundant sequences occur, the total number of redundant sequences for any peptide is small relative to the total size of the database; and(c) all protein sequences containing the peptide sequence that truly corresponds to the peptide sample used to generate the tandem spectrum have been removed.

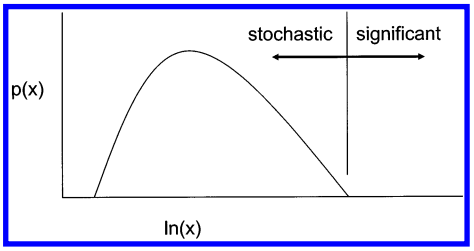

Figure 1. Schematic representation of a stochastic score distribution, as described in the text. Any peptide corresponding to a score within the body of the stochastic distribution cannot be confidently assigned as being a valid identification. A score higher than the right-hand boundary of the stochastic distribution may be assigned as potentially valid, with an associated expectation value.

The stochastic scoring distribution corresponding to a particular tandem spectrum on a particular database can then be measured using the following steps:

1. Find all peptides in the selected sequence database that match the experimental protein cleavage conditions (e.g., peptides generated by trypsin cleavage) and have parent ion masses within a selected value of d (eq 3).

2. Calculate the score for all of the peptides found in step 1 using the selected scoring system.

3. Record all of the scores generated in step 2.

4.Construct a histogram of the frequency of the occurrence of a particular score versus all of the scores collected in step 3.

Figure 1 illustrates the general form of the probability distribution calculated from the scoring frequency histogram using eq 4. It is then a simple matter to calculate the survival function, using eq 1. Some smoothing may be necessary to eliminate the effects of redundant sequences on the curves: alternately multiple identical scores caused by redundant sequences can be filtered out in step 3. Then eq 2 can be used to estimate how often a particular score would be expected to be found by random sequence matches. It is possible to assess how high a score generated from the spectrum and the valid peptide sequence should be to produce an assignment with a particular confidence level. Practically, it is very simple to construct the scoring frequency histogram "on-the-fly" as sequences are being scored, with very little computational overhead or impact on the overall performance of a practical scoring engine.

A much more interesting case is one in which the valid sequence for a particular peptide is not known in advance. The goal of this exercise is to determine the sequence of the peptide, rather than simply measure the stochastic distribution. In this case, condition c for the database cannot be met. If the same algorithm is applied in this case, conditions a and b on the database ensure that the majority of the scores recorded correspond to the stochastic distribution and that only a single scoring value (x^*) in the scoring frequency histogram will correspond to a valid match between the peptide and the tandem spectrum, regardless of how many times that peptide sequence occurs in the protein sequence database. A confident identification can be made if the valid peptide match generates a value for x^* in the "significant" portion of Figure 1. The actual confidence interval

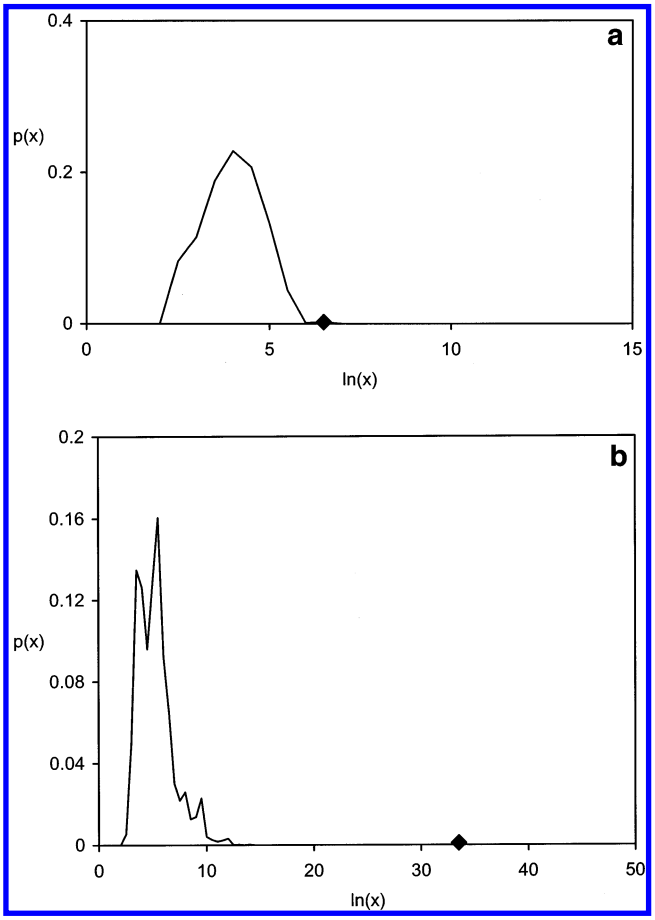

Figure 2. (a) Measured stochastic scoring histogram for a specific tandem mass spectrum, using the correlation-only scoring scheme, x_I (eq 6). The solid diamond marks the point in the histogram corresponding to the score for the valid peptide sequence (x^*). (b) Measured stochastic scoring histogram for the same tandem mass spectrum, using one of the alternate scoring systems, x_{II} (eq 7). The solid diamond marks the point in the histogram corresponds to the score for the valid peptide sequence (x^*).

of that identification can then be measured using the survival function and expectation value approach.

Figure 2 shows a real example of scoring histograms calculated using steps 1 to 4 above, a sample tandem spectrum, and the NCBI nr protein sequence database. The stochastic portion of the measured distributions is to the left of the histogram, and the single histogram point corresponding to the valid identified sequence (score x^*) is marked with a diamond. Clearly, x^* does not fall within the stochastic distribution in either case, indicating that it is unlikely that the match is the result of a stochastic match (see the previous section for a description of the spectrum and its valid sequence).

A few general comments can be made about the visual assessment of this type of histogram. The confidence that a score represents a valid identification increases as the distance between the high-scoring end of the stochastic distribution and x^* increases. In the case that poor spectral quality causes a valid match to fall within the stochastic distribution, there is simply insufficient information in the mass spectrum to confidently assign that mass spectrum to a particular peptide sequence. It should be noted that there is no way to distinguish between the case where there is insufficient data to positively match a sequence

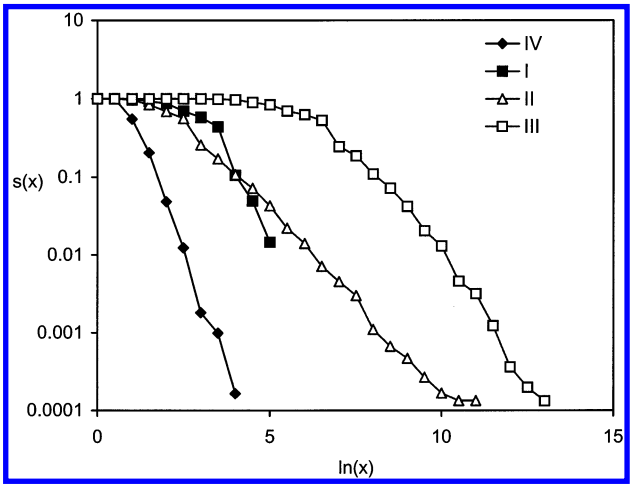

Figure 3. Measured survival functions for stochastic probability distributions using the four different scoring systems (eqs 6−9) described in the text.

with a spectrum and the case where the peptide sequence corresponding to the spectrum is not in the database (or has been modified in some unanticipated manner).

A much more sophisticated estimate of the confidence of a sequence assignment can be obtained from the survival function. A comparison of survival functions derived from the stochastic portions of the scoring histograms for each of the four scoring systems is illustrated in Figure 3, using the same spectrum, conditions, and database as for Figure 2. The arguments used to justify the use of the maximum formulation of the extreme value distribution[14] (also referred to as the Gumbel distribution) in sequence similarity scoring[11,12] are valid for these scoring systems and form the basis for the estimation of expectation values associated with survival functions. Based on this distribution, the high-scoring portion of a plot of $\log(s(x))$ versus $\log(x)$ should be linear, as is evident from the curves shown. Practically, a smoothed linear least-squares fit for all data points, $\log(s(x)) < \log(0.1)$, allows the estimation of the survival function—and hence the expectation value—for all x such that $s(x) > 0.1$. Applying this extrapolation process to the identification illustrated in Figure 2b, the expectation value for the point marked with a diamond is $e(x^*) = 9 \times 10^{-6}$. This expectation value can be interpreted as meaning that a stochastic match of that quality would only be expected to occur once in $\sim 10^5$ trials.

Inspection of Figure 2 may be used for the qualitative analysis of the differences between scoring systems. The value of x^*_{II} is clearly much better separated from the stochastic distribution than x^*_I. This enhanced relative separation means that this type of scoring has the potential to be more sensitive at detecting sequences in spectra with lower spectral quality. As indicated in the discussion above, system II scores rely on an additional step of spectrum interpretation, namely, the assignment of specific parent ion fragmentation reactions and the addition of that information to the score in an unbiased manner. The authors' experience in fully automated commercial and academic protein identification laboratories has shown this scoring system to be particularly useful for automating assignments in environments where it is not practical to have trained analysts refine protein identification lists manually. It should be noted, however, that the results of any scoring algorithm may vary depending on the quality

of the spectra used and the details of the parameters used for the including or excluding sequences from consideration by a particular search engine.

Measuring and fitting these distributions makes them practical for protein identification experiments, rather than using a time-consuming de novo calculation to derive them (e.g., a Monte Carlo simulation of the experiment). Based on our implementation of this method, measuring and fitting these distributions introduces a calculation overhead of <1%/spectrum. This measurement method also has the advantage of accurately determining the distribution appropriate for arbitrary combinations of search parameters, sequence databases, and mass spectra. This type of accuracy is difficult to obtain without unacceptable overhead from a theoretical calculation.

Figure 3 demonstrates how the survival functions can be used to normalize a scoring system, using the four example scoring systems described above. By extrapolation, a score of $\ln(x_I) = 10$ has an equivalent interpretation to $\ln(x_I) = 12$ and $\ln(x_{IV}) = 5$: the probability of that score or higher occurring at random is 0.01. Similarly, the interpretation of a score difference between $\ln(x_{IV}) = 5$ and 7 extrapolates to the difference between $\ln(x_{III}) = 20$ and 23: the probability that the score or higher will occur at random has decreased by 1 order of magnitude. This type of simple comparison based on a single graph has the additional advantage of uniquely taking into account the actual distribution of fragment ion intensities, the mass of the parent ion, the detailed makeup of the sequence database, and all anticipated sequence modifications that may have been allowed for in the original search.

A significant consideration in any protein identification experiment is the effect of changing the search parameters on the significance of the results. Intuitively, the smaller the allowed range for a parameter, the more significance should be placed on the best result. The degree to which this aphorism is true has been difficult to evaluate objectively using raw scores. Application of the survival function approach allows the quantification of the degree of benefit conferred by changes in any of the primary selection parameters. There are a large number of potential effects caused by changing these parameters, e.g., the number and identity of potential posttranslational modifications considered, the fragment ion mass error allowed, or the signal-to-noise ratio of the spectrum.

Rather than attempt an exhaustive survey of all possible effects, the effect of varying two such search parameters on the survival functions measured for a single tandem spectrum using scoring system II (eq 7) was selected as an example and is shown in Figure 4. The parent ion mass deviation d and the peptide cleavage specificity are permuted between two fixed values. The values $d = 2$ and the peptide cleavage specificity being trypsin (cleavage at lysine and arginine residues, except when followed by proline) represents values commonly using in protein identification experiments. The values $d = 100$ and no peptide cleavage specificity (cleavage at all residues) represent extreme conditions that are not commonly used. It has been the general perception of many in the field that such nonrestrictive parameter values have the potential to significantly affect the reliability of the results, masking valid matches. The use of nonrestrictive parameters also may dramatically increase the amount of time necessary for the identification calculation, e.g., an identification that uses nonspecific peptide cleavage will take ~10 times longer to calculate than one that that only considers tryptic peptides.

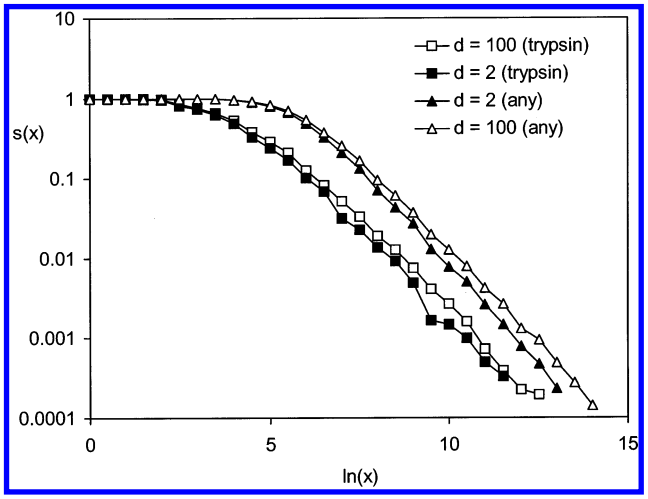

Figure 4. Survival functions measured for different search parameters, using scoring system x_{II} (eq 7). The symbol d represents the allowed deviation of the observed parent ion mass from the calculated mass of a peptide sequence (eq 3). The sequences were generated from intact protein sequences by assuming either cleavage with trypsin (lysine or arginine, except when followed by proline) or cleavage at any peptide bond.

Table 1. Comparison of Estimated Survival Function and Expectation Values for Valid Scores (x^*) Using Scoring Scheme x_{II} (Eq 7) in the Experimental Data Shown in Figure 4

parameters	$\log(s(x^*))$	$\log(e(x^*))$
$d = 2$ (trypsin)	−12.0	−4.6
$d = 100$ (trypsin)	−11.7	−4.6
$d = 2$ (any)	−11.5	−4.1
$d = 100$ (any)	−11.3	−3.8

Figure 4 demonstrates that varying these parameters has little effect on the overall form of $s(x)$. Comparing the results of one cleavage chemistry, varying d only results in a minor change in the slope of the plot. The difference in the survival functions for changing the cleavage specificity from trypsin to no specificity is more significant. The effect of altering the primary search parameters can be easily quantified by examining the values of the survival function and expectation value x^* (Table 1). It should be noted that $\ln(x^*) = 30.5$ for the valid score, which is unaffected by the choice made for d.

The survival function/expectation value calculations suggested above make drawing statistical inferences from large data sets routine exercises. As an example, take the task of analyzing a set of tandem mass spectra collected over the course of a liquid chromatography separation, with the goal of determining what protein sequences were represented in the original sample by identifying the peptide sequences from mass spectra. In this case, the confidence interval to calculate would be the confidence that the protein was present, rather than simply the confidence that individual peptides were present. A simple method to evaluate the expectation value for the protein, based on the individual expectation values for the peptides, requires only a slight refinement of the methods presented. The results of searching the appropriate sequence database with the complete set of tandem

mass spectra would be a series of potential peptide assignments, each with an associated expectation value. By setting a confidence limit threshold requiring that only peptides with $e(x^*) < e_{thres} < 1$ were to be considered, a subset of these assignments with the greatest likelihood of being valid would be created. By definition, such assignments would be rare and their stochastic occurrence would be governed by the Poisson distribution. Therefore, if there were M spectra collected during the analysis, the stochastic chances of $e(x^*) < e_{thres}$ is $M^{1/2}$ times greater than if a single measurement was made. Generalizing this relationship to multiple peptide assignments, the expectation value for the protein (E_{pro}) based on the experimental results would be given by the following expression

$$E_{pro} = \sqrt{M}\prod_{i=1}^{n} e_i(x_i^*) \tag{10}$$

where there are n valid peptide assignments contained in the protein sequence of interest, each with each peptide i having a characteristic expectation value and score.

CONCLUSIONS

The results of this paper suggest that the following steps may be used to reduce any well-behaved protein identification scoring system to an expectation value formulation for comparison with other systems.

1. Create the distribution $p(x)$ for each specific identification experiment, preferably experimentally by the methods described.
2. Calculate $s(x)$ from $p(x)$.
3. Fit the high-scoring portion of $s(x)$ to a model distribution.
4. Use the fitted function to extrapolate $s(x)$ to the desired score and convert the extrapolated value to $e(x)$.

The results shown and two years of practical experience using survival functions and expectation values to evaluate protein identification experiments have shown that this approach is valuable for automated, high-throughput experiments. They have also proved valuable in assessing the relative merit of scoring systems and evaluating the effects of altering primary identification parameters. We advocate the widespread use of these simple statistical measures to simplify and standardize the reporting of the confidence of protein identification results, allowing the users of different software to compare their results in a straightforward and statistically significant manner.

ACKNOWLEDGMENT

The authors thank the Manitoba Proteomics Centre at the University of Manitoba (http://www.proteome.ca) for their co-operation. They also thank Marcus Kalkum and Brian T. Chait of Rockefeller University for their discussions on this matter.

Received for review June 18, 2002. Accepted November 5, 2002.

AC0258709