research articles Journal of proteome research

Probity: A Protein Identification Algorithm with Accurate Assignment of the Statistical Significance of the Results

Jan Eriksson*,† and David Fenyö‡,§

Department of Chemistry, Swedish University of Agricultural Sciences, Box 7015, S-750 07, Uppsala, Sweden, Amersham Biosciences, 800 Centennial Avenue, Piscataway, New Jersey 08855, and The Rockefeller University, 1230 York Avenue, New York, New York 10021

Received July 4, 2003

An algorithm for protein identification based on mass spectrometric proteolytic peptide mapping and genome database searching is presented. The algorithm ranks database proteins based on direct calculation of the probability of random matching and assigns the statistical significance to each result. We investigate the performance of the algorithm by simulation and show that the algorithm responds to random data in the desired manner and that the statistical significance computed indicates the risk that a particular identification result is false.

Keywords: protein identification • algorithm • bioinformatics • mass spectrometry • proteomics • protein • peptide • peptide mapping • significance testing • simulation

Although the use of mass spectrometric protein identification has matured rapidly and become increasingly useful to the life sciences,[1−6] the development of accurate and well-documented methods for judging the quality of identification results and for judging the performance of identification algorithms is lagging behind. Here, we present Probity, a novel algorithm that provides a solution to the problem of accurately evaluating the quality of identification results. The algorithm ranks the proteins in a database by employing a direct computation of the risk that the matching between mass data and a protein in a database is random. We demonstrate that the correct assignment of the statistical risks that results are false can be computed directly and accurately. We also present general criteria for evaluating identification algorithms and demonstrate by simulation that the Probity algorithm handles the random matching in the desired way.

State-of-the-art proteome analysis often involves protein separation by 2D-gel electrophoresis.[7−10] The level of a protein is typically determined from the intensity of the spot on the gel. Each gel-spot of interest is excised and the protein is subjected to in-gel digestion using an enzyme that has high digestion specificity (e.g., trypsin). The resulting proteolytic peptides are extracted and analyzed by mass spectrometry (MS). One strategy for protein identification based on MS of a protein digest assumes that a set of proteolytic peptide masses provides a “fingerprint” of a particular protein. It is assumed that the peptide mass fingerprint can be recognized when searching a genome database.[7,8,11−15] Identification algorithms compute theoretical peptide masses assuming that each protein in the database is digested by the same enzyme as was used in the experiment. Subsequently, the algorithms calculate a score based on the number of matches between measured and calculated peptide masses. The score is used to rank the database proteins. In some algorithms, the score is simply the number of matches,[7,16] whereas in other algorithms, the score is the result of a computation based on the number of matches.[17,18] The protein (or proteins) with the best score is (are) identified. A common way of obtaining a further constraining protein “fingerprint” is to perform tandem mass spectrometry (MS/MS), whereby a single proteolytic peptide ion species is isolated and fragmented in the mass spectrometer.[19−22] Mass analysis of the resulting peptide fragment ions can yield highly constraining sequence information.

There is a risk of obtaining a *false* identification result, because each mass determined by MS has an error, $\pm\Delta m$, and can match several proteolytic peptides of various proteins in the database. We refer to the matches with proteins that are not present in the sample as random matches. An experiment will yield a false result when the score due to random matching is at least as good as the score of a real protein in the sample. A result is significant only if the experimental score deviates significantly from the scores that can be expected from false results. Testing of the significance of a result can be performed only if the score frequency function (distribution of scores) for random results has been established for the particular data and database search constraints of the experiment. We have previously demonstrated two methods to estimate frequency functions for random protein identification: (1) simulation, which includes repeated protein identification by searching a genome database and using different sets of random proteolytic peptide masses as data[23] and (2) model computation, in which the model always takes into account features of each individual set of peptide masses, genome database, and database search constraints.[24] Recently, a different approach was presented that estimates a frequency function based on statistics collected

[*] To whom correspondence should be addressed.
[†] Swedish University of Agricultural Sciences.
[‡] Amersham Biosciences.
[§] Rockefeller University.

10.1021/pr034048y CCC: $27.50
Published on Web 12/06/2003

during the search assuming that most proteins in the database to some degree randomly match an experimental set of proteolytic peptide masses.[22,25]

Here, we present a simplified approach utilizing an analytical formula for calculating the risk of random matching between experimental masses and theoretical masses of an individual database protein. The risk calculated is employed as a score for ranking the individual database proteins. The statistical significance of each result is subsequently calculated by appropriately including the frequencies of the various sizes of the proteins in the database. The accuracy of this approach is validated by simulation.

Materials and Methods

The *Saccharomyces cerevisiae* genome containing 6403 ORFs (NCBI, May 2000 release) was employed for studying the algorithm performance. Protein digestion was performed in silico assuming exposure to trypsin (cleaves with high specificity at the carboxyl side of lysine and arginine residues). Peptides within a mass region between 800 and 4500 Da were considered. The NCBI database was processed in order to obtain eight database subsets that cover different mass regions of the proteolytic peptides. Each database subset was sorted with respect to mass in order to enhance the search speed.

For the validation of the Probity algorithm, random sets of proteolytic peptide masses were used. These sets were generated by choosing each mass randomly from the theoretical proteolytic peptides of different randomly chosen proteins in the database. Each protein in the database was allowed to contribute with at most one proteolytic peptide mass to an individual random set.

Scripts written in *Perl* and programs coded in *C* were employed for the computations, which were performed on a Dell Optiplex GX1 (550 MHz Pentium III) personal computer.

Algorithm. Earlier, we presented a model that describes accurately the process of random matching of measured and theoretical peptide masses.[24] We showed that the application of this model can determine directly the significance of an identification result when the proteins are *ranked strictly by the number of matches*. Here, we show that the model computation of the probabilities for random matching can be employed directly in the *ranking* of the proteins in the database.

The model of random matching employs detailed features of the distribution of peptide masses. Proteolytic peptide masses are distributed in discrete clusters or peaks due to the almost integral values of the masses of the atoms (C, H, N, O, S) from which the peptides are composed. We refer to these clusters as peptide *mass distribution peaks*. For a mass accuracy, Δm, better than ±0.25 Da, the random matching always occurs within a single mass distribution peak.

The algorithm detects the number of matches, k, between the experimental set of peptide masses and the theoretical peptide masses of each protein in the database. Assuming a data set with n masses, a massaccuracy of Δm, and a maximum number of missed cleavage sites u, the probability for an individual protein having k_u proteolytic peptides to yield k matches by chance is given by eq 1

$$p(k) = \sum_{k_i, \sum k_i = k} \left\{ \prod_{i=1}^{q} \binom{n_i}{k_i} \cdot p_i'^{ki} \cdot (1 - p_i')^{n_i - k_i} \right\} \qquad (1)$$

The probabilities p'_i of eq 1 are given by eq 2

$$p_i' = f_i \cdot \frac{k_u}{m_{i+1} - m_i} \cdot \delta(i, \Delta m, u) \qquad (2)$$

The index i denotes a mass region for which the heights and widths of the mass distribution peaks are assumed to be constant. Eight different mass regions are used here ($q = 8$ in eq 1) to cover the entire mass range (800−4500 Da). The influence of q on the computational accuracy has been examined in ref 24. The symbol n_i denotes the number of proteolytic peptide masses experimentally observed in mass region i, $m_{i+1} - m_i$ denotes the number of mass distribution peaks in region i, and f_i denotes the fraction of the total number of peptides that are estimated to belong to region i in the database. We assume that the fraction f_i of the proteolytic peptides from an individual protein that on the average falls into i can be estimated from the fraction of the total number of proteins in the database yielding peptides within i. $\delta(i, \Delta m, u)$ denotes a function that depends on the shape of the peptide mass distribution peak. A detailed description on how to determine $\delta(i, \Delta m, u)$ is given in ref 24. The probability that an individual protein having k_u proteolytic peptides will yield *at least k′* matches by chance is

$$\beta = 1 - \sum_{k < k'} p(k) \qquad (3)$$

Below, we refer to $-\log(\beta)$ as the *score*. An ensemble of different *random* sets of peptide masses, where each set has the same number of members and for which the same search constraints Δm and u are employed, will yield a distribution of different β-values for the proteins identified. Therefore, β can be analyzed statistically. From this distribution it is possible to estimate the probability of observing a value $\leq \beta$ by chance.

Assuming that any of the proteins in the database could yield $\leq \beta$ and assuming that k is limited by n and k_u, the probability, S, of observing $\leq \beta$ by chance is

$$S = P(\text{at least 1 protein will yield} \leq \beta) =$$

$$1 - P(\text{all proteins yield} > \beta) = 1 - \prod_{k_u=1}^{k_u^{\max}} P_{k_u}(> \beta) \qquad (4)$$

with

$$P_{k_u}(>\beta) = \left\{ \sum_{k=0}^{k=k_{\max}} p(k) \right\}^{\Psi_{k_u}} \qquad (5)$$

where $k_{\max}$ is determined by $\beta$and ψ_{ku} is the frequency of a particular k_u value in the database and $p(k)$ is given by eq 1. Hence, S is the value that indicates the statistical significance[26] of the protein identification result when β is used to rank the proteins in the database.

Results

Response to Random Data. Others and we have noted that protein identification algorithms that rank strictly based on the number of matches favor large proteins[24] (Figure 1). Hence, when the data quality is poor, false high-mass proteins are typically identified. An algorithm should not favor any particular protein size. We investigated the potential favoring of protein sizes by employing random sets of proteolytic peptide masses, where each mass had been randomly chosen from the theoretical proteolytic peptides of a randomly chosen protein in the database. It has been shown previously that the value

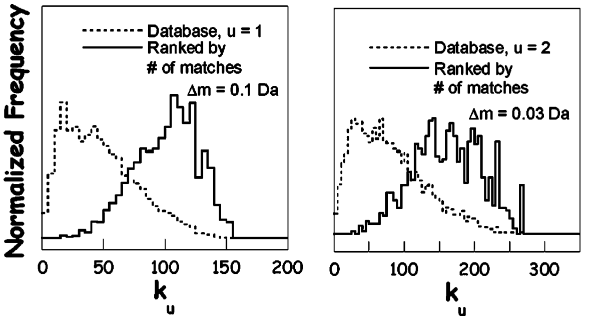

Figure 1. Size distribution of randomly identified proteins using ranking strictly by the number of matches compared with the size distribution of proteins in the database for various random peptide mass data and different search constraints (20 peptides, *n*, maximum number of missed cleavage sites, *u*, and mass accuracy, Δ*m*).

k_u is an appropriate measure of protein size.[24] If many sets of random proteolytic peptide masses are employed for searching a database with a particular algorithm, then it is possible to obtain a distribution of the k_u values of the proteins randomly identified. The distribution of k_u values of randomly identified proteins, henceforth response distribution, indicates how the algorithm responds to random (or poor quality) data. If no particular protein size is favored, then the response distribution is close to the distribution of the sizes of the proteins in the database. Figure 1 displays clearly how ranking strictly by the number of matches strongly favors the random identification of large (high k_u value) proteins. Figure 2 displays the superior performance of the Probity algorithm, which yields response distributions for all search constraints that closely resemble the size (k_u) distributions of the database proteins.

Score and Significance Relationship Determined by Simulation. Simulation by repeated database searching with different sets of random peptide masses yield a distribution of scores for random identifications (frequency function).[23] In general, the protein identification algorithms yield a score distribution for randomly identified proteins that is strongly dependent on the peptide mass data and the search constraints.[23] In contrast, Probity yields score (eq 3) distributions that are highly similar for different search constraints (Figures 3 and 4). This is due to the precise description of the random matching process by the algorithm, leading to an intrinsic scaling of the score between different search constraints. Figure 4 displays the significance derived from the various distributions displayed in Figure 3 by constructing the cumulative relative score frequency beginning with the best score.[23] Hence, the diagram of Figure 4 represents the relationship between the score and the significance for Probity.

Score and Significance Relationship Determined by Computation. A collection of different random sets of peptide masses, where each set has the same number of members and for which the same search constraints Δ*m* and *u* are employed, will yield a distribution of different β-values for the protein identified (Figure 3). Hence, the relative frequency of β with repeated identifications under the same constraints but with different random sets of peptide masses is the true indicator of the statistical significance (Figure 4). To verify that our algorithm correctly predicts the statistical significance of identification results, we performed simulations of protein identifications using different random sets of peptide masses. It is seen in Figure 5 that the computed significance values and the simulated relative frequencies for the corresponding individual scores (β) agree very well.

Discussion

Computation versus Simulation. The good agreement between the significance values obtained by direct computation (eq 4) and the significance values derived by simulation (Figure 5) indicates that the simple and straightforward computational approach of the Probity algorithm is robust. It is evident that our computational approach provides a link between the random matching probability of an individual protein and the global statistical significance taking into account characteristic features of the database.

The score (eq 3) as well as the significance computed are continuous variables (probabilities). Nondiscrete variables are generally desirable in ranking processes. Sometimes an experimentalist desires a discrete method in the sense that results that do not display a better significance than a predefined level, *S*′, are discarded. This significance testing approach is generally possible when the frequency function (score distribution for random results) for the particular experiment is known. Hence, an alternative use of the Probity algorithm is to employ the inverse of eq 4 (Figure 4) to derive the critical score, β_C required to obtain the significance, *S*′, desired by the experimentalist.

Inclusion of Additional Information. We are currently refining some of the features of the Probity algorithm, including investigating the possibilities of utilizing the characteristics of the Probity algorithm to automate partially the process of

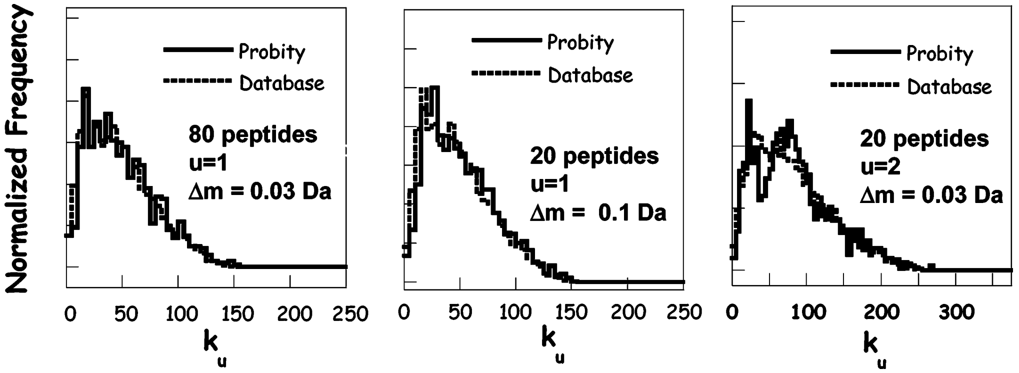

Figure 2. Size distribution of randomly identified proteins using Probity compared with the size distribution of proteins in the database for various random peptide mass data and different search constraints (number of peptides, *n*, maximum number of missed cleavage sites, u, and mass accuracy, Δ*m*).

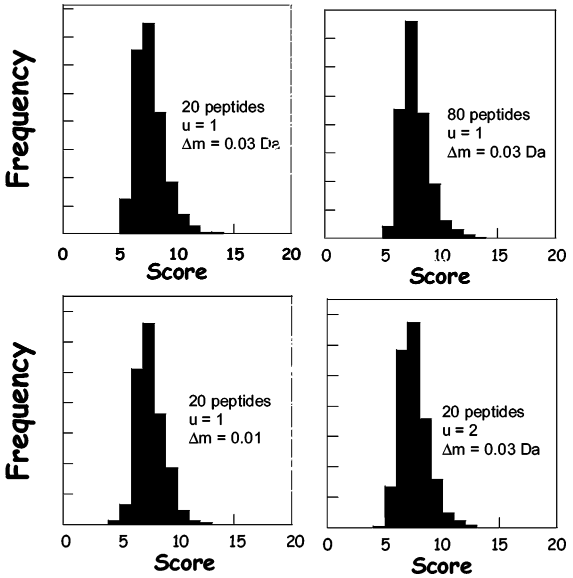

Figure 3. Simulated score distribution of randomly identified proteins using Probity for various random peptide mass data and different search constraints (number of peptides, *n*, maximum number of missed cleavage sites, *u*, and mass accuracy, Δ*m*).

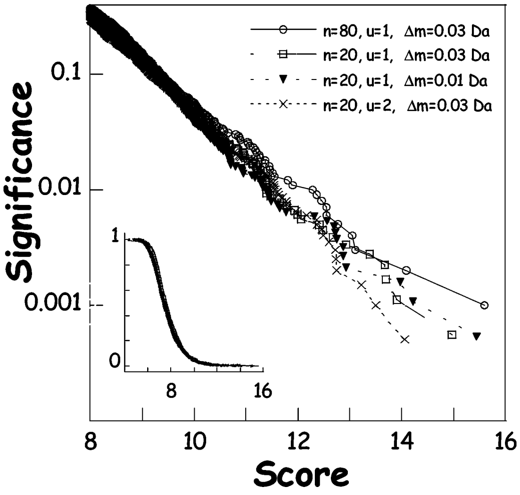

Figure 4. Simulated relationship between significance and score using Probity for various random peptide mass data and different search constraints (number of peptides, *n*, maximum number of missed cleavage sites, *u*, and mass accuracy, Δ*m*). *The inset displays all data points on linear scales.*

choosing database search constraints. Other future improvements will include deriving continuous functions that can approximate the discrete functions employed here. This will facilitate the inclusion of weight functions based on peak intensity information and deviations between theoretical and experimental mass values as well as MS/MS data taking into account fragmentation systematic.

Automated Applications. Once our algorithm is interfaced with a user environment, Probity can provide a highly useful tool for the protein identification practitioner, by optimizing the use of the experimental information. We believe that the features of the Probity algorithm shown here could become very useful to *automated* applications.[27,28] The direct and accurate computation of significance provided by Probity is a desirable feature in automated systems, especially if the information can be transferred in real-time to the mass spectrometer acquiring the data. The Probity algorithm handles random matching accurately and, therefore, has the potential to improve accurate protein identification in complex protein *mixtures*, where the mass information from each protein can be viewed as a random background to the mass information from each of the other proteins in the mixture.

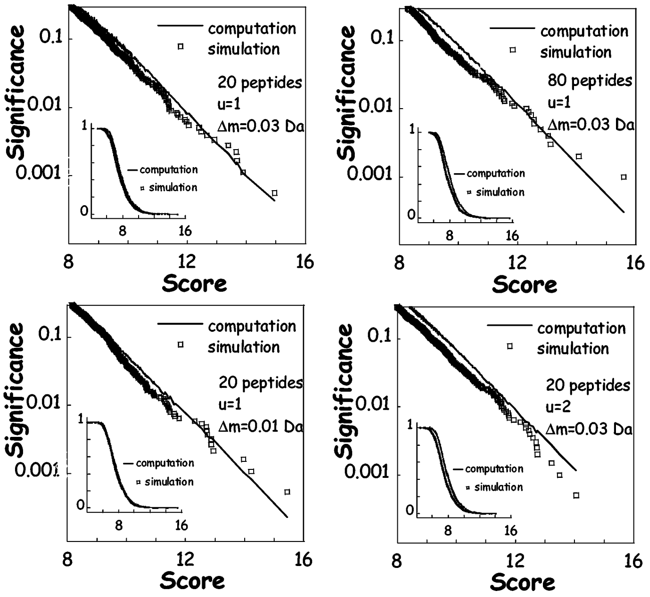

Figure 5. Comparison of computed and simulated statistical significance as a function of the score for various random peptide mass data and different search constraints (number of peptides, *n*, maximum number of missed cleavage sites, *u*, and mass accuracy, Δ*m*). *The inset displays all data points on linear scales.*

Conclusions

The algorithm presented addresses the two central problems in protein identification: (i) the optimal use of the experimental information to allow for identification of low abundance proteins, and (ii) the accurate assignment of the probability that a result is a false positive. We have shown here that Probity responds to random data in the random manner desired. We have also shown that a necessary condition for accurately computing the statistical risk (significance level) that a protein identification result is false is to appropriately include in the computations the frequencies of the various sizes of the proteins in the database. Simulations employing random mass data verified the accuracy of the computed significance level associated with each result.

Acknowledgment. This work was supported by grants to J.E. from the Swedish Research Council for Forestry and Agriculture and from the Swedish Research Council.

References

(1) Aebersold, R.; Mann, M. *Nature* **2003**, *422*, 198.
(2) Gavin, A. C. et al. *Nature* **2002**, *415*, 141.
(3) Ho, Y. et al. *Nature* **2002**, *415*, 180.
(4) Andersen, J. S.; Mann, M. *FEBS Lett.* **2000**, *480*, 25.
(5) Hanash, S. *Nature* **2003**, *422*, 226.
(6) Anderson, N. L.; Anderson, N. G. *Mol. Cell Proteomics* **2002**, *1*, 845.
(7) Henzel, W. J.; Billeci, T. M.; Stults, J. T.; Wong, S. C.; Grimley, C.; Watanabe, C. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90*, 5011.
(8) Shevchenko, A. et al. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93*, 14 440.
(9) Jensen, O. N.; Wilm, M.; Shevchenko, A.; Mann, M. *Methods Mol. Biol.* **1999**, *112*, 513.

(10) Fey, S. J.; Larsen, P. M. *Curr. Opin. Chem. Biol.* **2001**, *5*, 26.
(11) Jensen, O. N.; Larsen, M. R.; Roepstorff, P. *Proteins* **1998**, *Suppl*, 74.
(12) Mann, M.; Wilm, M. *Anal. Chem.* **1994**, *66*, 4390.
(13) Mortz, E.; Vorm, O.; Mann, M.; Roepstorff, P. *Biol. Mass Spectrom.* **1994**, *23*, 249.
(14) James, P.; Quadroni, M.; Carafoli, E.; Gonnet, G. *Biochem. Biophys. Res. Commun.* **1993**, *195*, 58.
(15) Beavis, R.; Fenyo, D. *Proteomics: A Trends Guide* **2000**, 22.
(16) Mann, M.; Hojrup, P.; Roepstorff, P. *Biol. Mass Spectrom.* **1993**, *22*, 338.
(17) Pappin, D. J. C.; Hojrup, P.; Bleasby, A. *Curr. Biol.* **1993**, *3*, 327.
(18) Zhang, W.; Chait, B. T. *Anal. Chem* **2000**, *72*, 2482.
(19) Yates, J. R. D.; Eng, J. K.; McCormack, A. L.; Schieltz, D. *Anal. Chem.* **1995**, *67*, 1426.
(20) Haynes, P. A.; Fripp, N.; Aebersold, R. *Electrophoresis* **1998**, *19*, 939.
(21) McLafferty, F. W.; Kelleher, N. L.; Begley, T. P.; Fridriksson, E. K.; Zubarev, R. A.; Horn, D. M. *Curr. Opin. Chem. Biol.* **1998**, *2*, 571.
(22) Field, H. I.; Fenyo, D.; Beavis, R. C. *Proteomics* **2002**, *2*, 36.
(23) Eriksson, J.; Chait, B. T.; Fenyo, D. *Anal. Chem* **2000**, *72*, 999.
(24) Eriksson, J.; Fenyo, D. *Proteomics* **2002**, *2*, 262.
(25) Fenyo, D.; Beavis, R. C. *Anal. Chem.* **2003**, *75*, 768.
(26) Davies, O. L.; Goldsmith, P. L. *Statistical Methods in Research and Production*; Longman Group Ltd: London, 1976.
(27) Jensen, O. N.; Mortensen, P.; Vorm, O.; Mann, M. *Anal. Chem.* **1997**, *69*, 1706.
(28) Gras, R. et al. *Electrophoresis* **1999**, *20*, 3535.

PR034048Y