Jan Eriksson[1]
David Fenyö[2]

[1]Department of Chemistry,
Swedish University of
Agricultural Sciences,
Uppsala, Sweden
[2]Proteometrics LLC,
New York, USA

A model of random mass-matching and its use for automated significance testing in mass spectrometric proteome analysis

A rapid and accurate method for testing the significance of protein identities determined by mass spectrometric analysis of protein digests and genome database searching is presented. The method is based on direct computation using a statistical model of the random matching of measured and theoretical proteolytic peptide masses. Protein identification algorithms typically rank the proteins of a genome database according to a score based on the number of matches between the masses obtained by mass spectrometry analysis and the theoretical proteolytic peptide masses of a database protein. The random matching of experimental and theoretical masses can cause false results. A result is significant only if the score characterizing the result deviates significantly from the score expected from a false result. A distribution of the score (number of matches) for random (false) results is computed directly from our model of the random matching, which allows significance testing under any experimental and database search constraints. In order to mimic protein identification data quality in large-scale proteome projects, low-to-high quality proteolytic peptide mass data were generated *in silico* and subsequently submitted to a database search program designed to include significance testing based on direct computation. This simulation procedure demonstrates the usefulness of direct significance testing for automatically screening for samples that must be subjected to peptide sequence analysis by *e.g.* tandem mass spectrometry in order to determine the protein identity.

Keywords: Protein identification / Proteome / Database / DNA sequence / Bioinformatics / Mass spectrometry / Significance / Automation Pro 0177

1 Introduction

Proteome projects are expected to emerge [1] in the wake of completed genome projects [2–6]. State of the art proteome analysis typically involves protein separation by 2-D gel electrophoresis followed by protein identification based on mass spectrometry (MS) peptide mapping and genome database searching [7–9]. High throughput and extensive automation [10, 11] are highly desirable features of systems for proteome analysis. Robot-handled gel spot excision, in-gel protein digestion, and sample preparation for MS analysis can follow the rapid computerized read-out of protein expression levels on a gel. Automated assessment of the quality of each identification result will become critical in any system for proteome analysis with limited human intervention. We present here a rapid computational tool for solving the problem of automated quality assessment of protein identification results obtained by MS analysis of protein digests. The tool performs a significance test and is based on a model that describes in detail the phenomenon of random matching that can lead to false identification results.

Protein identification based on MS of a protein digest [12–18] assumes that a pattern of proteolytic peptide masses provides a “fingerprint” of a particular protein and that the fingerprint can be recognized when searching a genome database. The protein digestion is usually done with a proteolytic enzyme having high digestion specificity (*e.g.* trypsin). Identification algorithms compute the number of matches between peptide masses from the experiment and the peptide masses from individual proteins in a database, assuming that each protein in the database is digested by the same enzyme as was used in the experiment. A score is used to rank the database proteins. In some algorithms, the score is simply the number of matches [19], whereas in other algorithms the score is the result of a computation based on the number of matches [13, 20]. The protein or proteins with the best score is (are) identified. 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

Correspondence: Dr. David Fenyö, Proteometrics LLC, 7W 36th Street, New York, NY 10018, USA
E-mail: fenyo@proteomics.com

the matches with proteins that are not present in the sample as random matches. A modified peptide will yield random matches only. A false result is obtained 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 demonstrated previously the method of using simulations to estimate frequency functions, *f(S)*, for random protein identification [21]. That method involves two steps: (1) generation of many different random proteolytic peptide maps from a genome; and (2) simulation of protein identification by searching a genome database and using the random proteolytic peptide maps as data. A simulation with a set of different random peptide maps with the same number of masses yields *f(S)* for random protein identifications characteristic for that peptide map size and other constraints used in the database search.

Here, we derive a model of the random matching of measured and theoretical peptide masses and employ this model to compute score frequency functions, *f(S)*, for random protein identifications when using an algorithm that ranks proteins according to their number of matches. The model computations elucidate the nature of the process of random matching and always take into account features of each individual peptide map, genome database, and database search constraints. The method presented here provides a much faster and also a more accurate means of computing *f(S)* compared with the method of simulation. We demonstrate that the rapid and accurate model computation provides a useful tool for automated testing of the significance of protein identification results.

2 Materials

Three different genome databases were employed: *Haemophilus influenzae*, *Saccharomyces cerevisiae*, and *Caenorhabditis elegans*, containing respectively 1718, 6403 and 19 100 (May 2000 release) ORFs. Protein digestion was performed *in silico* assuming exposure to trypsin (trypsin cleaves with high specificity at the carboxyl side of lysine and arginine residues). Only tryptic peptides with a mass between 800 Da and 4500 Da were considered. Scripts written in Perl were employed for all computations, which were performed on Dell Optiplex GX1 (550 MHz Pentium III) or Dell Precision 210 (500 MHz Pentium III) personal computers.

2.1 Model

A model that allows direct computation of frequency functions describing the scores (number of matches) of randomly identified proteins must describe accurately the process of random matching of measured and theoretical peptide masses. Therefore, a model must be designed to take into account specific information about the database proteins and their proteolytic peptides.

2.1.1 Protein size

Others and we have noted that identification algorithms that rank proteins according to their number of matches tend to favor large proteins [13, 21 and Roepstorff, personal communication]. Figure 1 (left, top panel) displays the mass distribution of the proteins in a genome database compared with the mass distribution of the proteins identified in simulations using random proteolytic peptide maps and ranking by the number of matches. The preference for random identification of high mass proteins is clear, which indicates that the protein size is a key quantity in the description of random matching. The protein mass, M_p, correlates with the theoretical number, k_u, of proteolytic peptides that a protein in a database can yield (Fig. 1, right, top panel). The spread of k_u values around a given value of M_p reflects the influence of the protein sequence (*e.g.*, some membrane proteins yield very few tryptic peptides). The distribution of k_u-values of the proteins in a database depends on the maximum number, *u*, of missed cleavage sites assumed (Fig. 1, left, bottom panel). It is seen in Fig. 1 (right, bottom panel) that proteins that can yield many proteolytic peptides dominate the random matching. In order to model the sequence-dependence appropriately, we use the quantity k_u in the description of random matching.

2.1.2 Peptide mass distribution peaks

Proteolytic peptide masses in a genome database 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 [22]. We will henceforth refer to these clusters as peptide mass distribution peaks. The widths of the peptide mass distribution peaks increase with increasing mass (Fig. 2). For unmodified peptides of mass < 4500 Da determined to an accuracy better than $\pm$ 0.25 Da, the random matching always occurs within a single mass distribution peak.

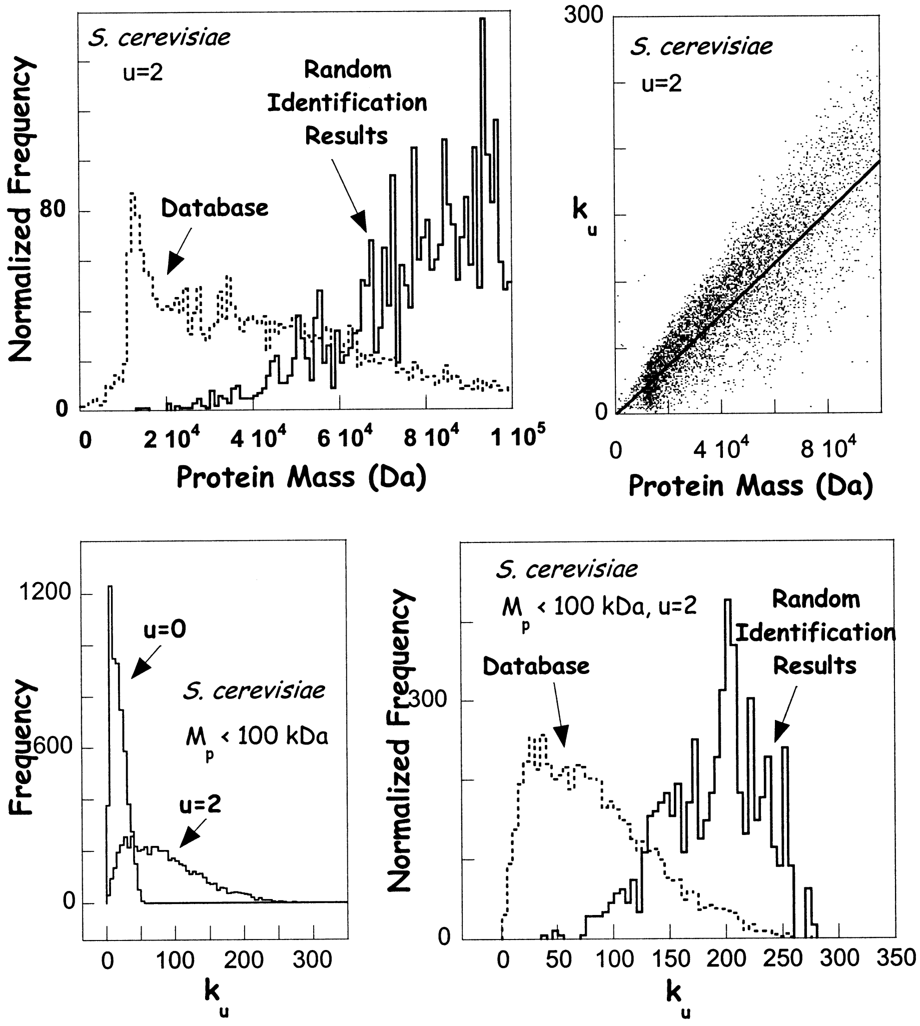

Figure 1.

2.1.3 Peptide mass frequencies

The frequency of proteolytic peptides in a genome database decreases with increasing mass. Hence, the higher the mass the lower the number of proteins in the database that can match randomly a measured proteolytic peptide mass within the mass accuracy ±Δm (Fig. 3).

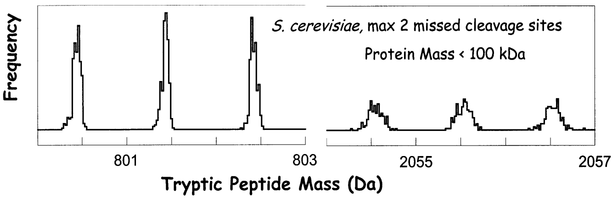

Figure 2.

2.1.4 Mass regions

The frequency of proteolytic peptides within the error of a mass measurement varies with mass as illustrated by Figs. 2 and 3. The variable frequency hinders a simple statistical description of random matching that directly covers the entire mass range of proteolytic peptides. We approached this problem by assuming that the mass range can be divided into a finite number, q, of proteolytic peptide mass regions and that within each region, i, frequencies of proteolytic peptides are approximately constant. We can thereafter obtain a statistical description of the random matching over the entire mass range by combining appropriately the probabilities of the number of random matches computed for each mass region.

2.1.5 Probabilities

A defined proteolytic peptide mass region, i, contains $(m_{i+1} - m_i)$ mass distribution peaks between the masses m_i and m_{i+1}. We assume that the fraction f_i of the proteolytic peptides from an individual protein that on the aver-

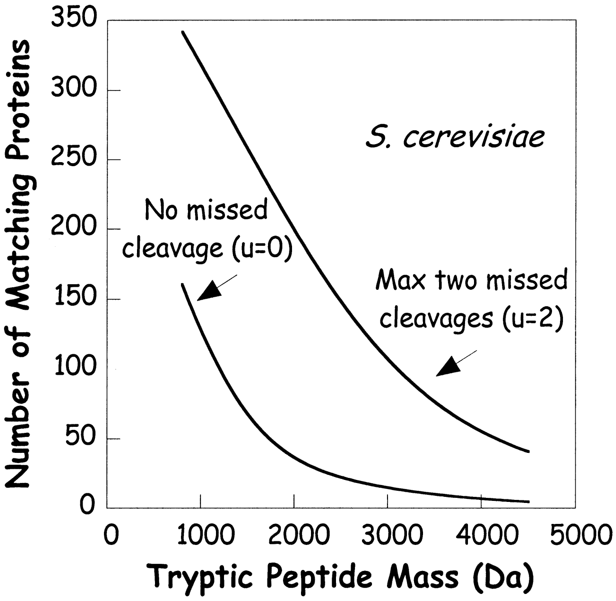

Figure 3.

age falls into *i* can be estimated from the fraction of the total number of proteins in the database yielding peptides within *i* (Fig. 3). Knowing f_i, the probability, p_i, that a proteolytic peptide from a particular protein characterized by k_u will be found in a single randomly chosen peptide mass distribution peak in region *i* can be computed as:

$$p_i = f_i \frac{k_u}{m_{i++} - m_i} \tag{1}$$

The probability, p'_i, of finding a proteolytic peptide originating from a particular protein characterized by k_u within a region $\pm\Delta m$ around a randomly chosen proteolytic peptide mass *m* is:

$$p'_i = p_i \delta\,(i, \Delta m) \tag{2}$$

where δ (*i*, Δm) denotes a function that depends on the shape of the (peptide) mass distribution peak and *i* refers to a (peptide) mass region δ (*i*, Δm) can be interpreted as a statistical measure of the fraction of the proteolytic peptide masses of a peptide mass distribution peak that can be found within $\pm\Delta m$ from a randomly chosen peptide mass. The derivation of δ (*i*, Δm) is described in the Appendix.

The probabilities, p_i', computed for all mass regions, *i*, can be employed to compute a total probability, *p(k)*, for an individual protein in the database to match randomly *k* out of *n* masses, where the *n* masses refers to the number of proteolytic peptide masses in the mass data (proteolytic peptide mass map).

$$p(k) = \sum_{k_i, \sum k_i = k} \left\{ \frac{n_1!}{k_1!(n_1 - k_1)!} p_1'^{k_1} (1 - p_1')^{n_1 - k_1} \right.$$

$$\frac{n_2!}{k_2!(n_2 - k_2)!} p_2'^{k_2} (1 - p_2')^{n_2 - k_2} \ldots \tag{3}$$

$$\left. \frac{n_a!}{k_q!(n_q - k_q)!} p_q'^{k_q} \left(1 - p_q'\right)^{n_q - k_q} \right\}$$

where *q* denotes the number of mass regions, n_1 denotes the number of masses in the mass data that are in mass region 1, n_2 denotes the number of masses in the mass data that are in mass region 2 *etc.*, and k_i, where *i* = 1, 2, . . ., *q*, denotes the number of matches in mass region *i*. The values of k_i are all combinations of values that apply to the constraint $\sum_i k_i = k$.

2.1.6 Frequency functions

The knowledge of *p(k)* for a particular experimental condition, provides a method of generating a frequency function of scores (number of matches) for random identifications of proteins. Since the random matching is dominated by proteins that have large k_u values (Fig. 1), we can simplify the computation by using a subpopulation of database proteins having large k_u values. The frequency function is denoted *f(S)*, where *S* is a score (number of random matches). With *S* = *k*′,

$$f(S) = \left\{ \sum_{kk=0}^{k'} p(k) \right\}^H - \left\{ \sum_{k=0}^{k'-1} p(k) \right\}^H \tag{4}$$

where *H* is the number of proteins in the subpopulation.

The formal derivation of Eq. (4) is given in the appendix. We found by inspection of results from simulations of random protein identification and by iterated use of Eq. (4) that a good agreement between simulated and computed *f(S)* is obtained if the size, *H*, of the subpopulation is equal to the number of proteins in the database having k_u values corresponding to at least 70% of the maximum k_u value in the database (for a particular number of allowed missed cleavage sites and a particular assumed maximum protein mass) and if the *median* k_u value ($0.85 \cdot k_u^{max}$) in the subpopulation is used as an approximate representative k_u value for all the *H* members of the subpopulation.

2.1.7 Critical score and significance testing

The frequency function, *f(S)*, for random identification results is the basis for testing the significance of an experimental identification result characterized by a score

S_E. Employing S_E as the test variable, the hypothesis H_0: *"the result is random"*, is rejected at the *significance level* α, if $S_E \geq S_C$. S_C is referred to as the critical score and is derived from the relation:

$$\sum_{S \geq S_C} f(S) \leq \alpha \qquad (5)$$

α is chosen prior to the significance test [23] and represents the test error risk or the statistical risk that the hypothesis H_0 is rejected although it is actually correct. α should be small and often either of the values 5%, 1%, or 0.1% are used [24]. *f(S)* or S_C must be known for the particular conditions of a given experiment, since their magnitudes depend on all pertinent experimental constraints (number of mass peaks, mass accuracy, *etc.*).

3 Results

3.1 Comparison of computed and simulated frequency functions and critical scores

Each frequency function obtained by using Eq. (4) is data dependent, since the probabilities computed depend on the numbers n_i describing how many of the n masses in a map that fall into each respective mass region i. In contrast, the simulation approach typically yields a frequency function based on all the different (typically $>$1000) random tryptic peptide maps employed [21]. In order to compare the frequency functions and critical scores generated by Eq. (4) with those generated by simulations, we computed averages of probabilities computed when using $>$100 different distributions of n_i, where each value of n_i was randomly chosen under the constraint $\Sigma n_i = n$. Results obtained by simulations and results obtained by averaging over a large set of computations agree well (Fig. 4).

The use of a simulated frequency function as a basis for significance testing will yield an accurate result as long as the experimental peptide map has approximately the same mass distribution as that in the whole database (Fig. 3). However, the mass distribution of individual peptide maps can sometimes deviate strongly from the expected overall mass distribution. Figure 5 shows simulated and computed frequency functions for three different cases, where in two cases the masses in the maps have been restricted and differ from the average mass distribution. In case (1) all the masses in the maps (each with 20 peptides) are between 1396 Da and 4500 Da. In case (2) the entire mass range 800 Da to 4500 Da is used and in case (3) all the masses are between 800 Da and 1396 Da. The frequency functions for the three cases display clear differences. For example, the critical scores corresponding to the 0.1% significance level are in the cases (1), (2) and (3): 8, 10 and 11 matches respectively. To perform simulations for every possible distribution of masses in an individual peptide map in addition to all other variable constraints is time consuming, whereas the model computation takes the peptide mass distribution as well as all other constraints into account in a direct and rapid way.

3.2 Automated significance testing

In order to demonstrate the concept of direct automated significance testing, the model-based testing of significance was directly implemented into a protein identification algorithm that ranks proteins according to their number of matches. The performance of the automated significance testing was investigated by simulating a set of protein identifications, where in each map a fraction of the masses originated from a single randomly chosen protein (correlated masses) and the rest of the masses were each from a different protein (noncorrelated). In each map, a different randomly chosen protein was employed to generate the correlated masses, which were randomly chosen and corresponded to a randomly chosen protein sequence coverage in the range 15–65%. Figure 6 displays the results for protein identification based on maps with 35 peptide masses for the organism *S. cerevisiae*. It is seen in Fig. 6 that the significance testing efficiently rejects false results. However, also a fraction of the true results are discarded by the significance test. As a lower frequency of false results is tolerated, more true results become nonsignificant.

4 Discussion

4.1 Discarded results

In an automated system for proteome analysis, the results that do not pass the significance test must be analyzed further. A common way of obtaining further constraining information is to perform tandem mass spectrometry (MS/MS), whereby a single ion species is isolated and fragmented in the mass spectrometer [24–26]. Mass analysis of the resulting fragment ions can yield highly constraining sequence information. MS/MS is more time consuming than is peptide mass mapping. We envision that in order to maximize throughput, sensitivity and quality, the automated significance testing can direct the proteome analysis system to the samples that need MS/MS analysis.

4.2 Computational accuracy and speed

We have stated above in Section 3.1 that the model based computation of the frequency function needed for significance testing is more accurate than that resulting

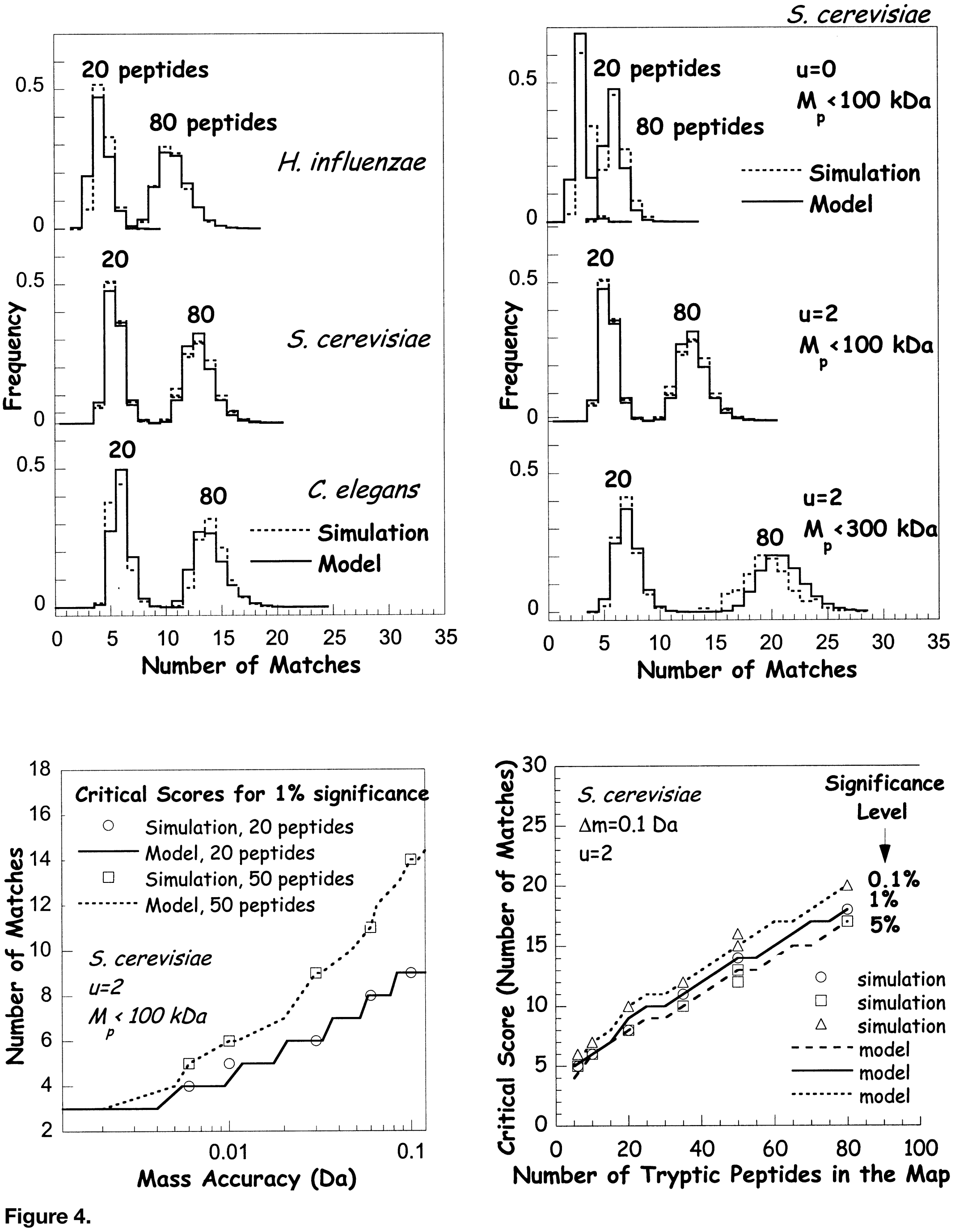

Figure 4.

from a simulation, provided that simulations are not performed for any given distribution of masses in a peptide map. In a practical application, not only the accuracy is important, but also the time it takes to obtain the result. We will therefore discuss briefly what is the computation time and what pertinent factors are influencing the computation time. Frequency functions are computed by solving the Eq. 3 and 4. In the computations of frequency functions and critical scores shown in Figs. 4 and 5 respectively, we divided the tryptic peptide mass range (800 Da–4500 Da) into four mass regions. Figure 7 displays a comparison between the use of 2, 4 and 8 mass regions in the model computations. It is seen that the agreement between simulation and computation improves with an increasing number of mass regions. However, the agreement is very good already at the use of four mass regions. The computations become slower with an increasing number of mass regions due to the

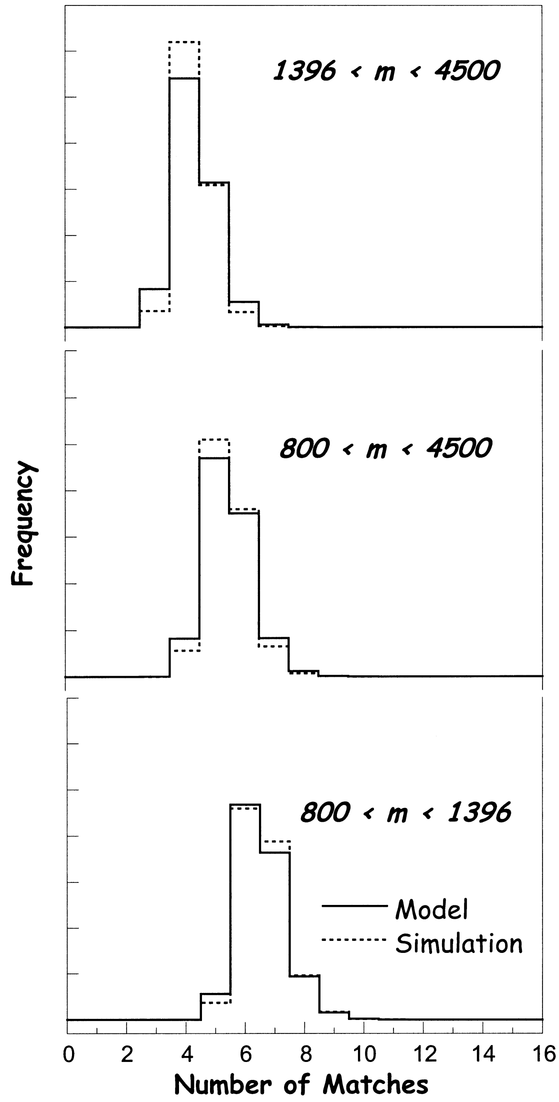

Figure 5.

increasing number of terms in Eq. (3). With our current software and computers, a single frequency function (which is all one needs in a practical application) takes 0.003 to 4 s to generate (n = 5–80) by direct computation when using four mass regions, whereas with eight mass regions the corresponding numbers are 0.008 to 10 s. It takes about 1000 s to derive a frequency function by simulation (using a program written in C [21]). The Perl scripts employed here for testing the model computation concept are not optimized with respect to computational speed. The general use of more than four mass regions in the computations can be facilitated by implementing the model in a compiled code such as C, which would according to our experience speed up the computations by a factor ≫10, hence yielding computation times ≪1 s for any significance test.

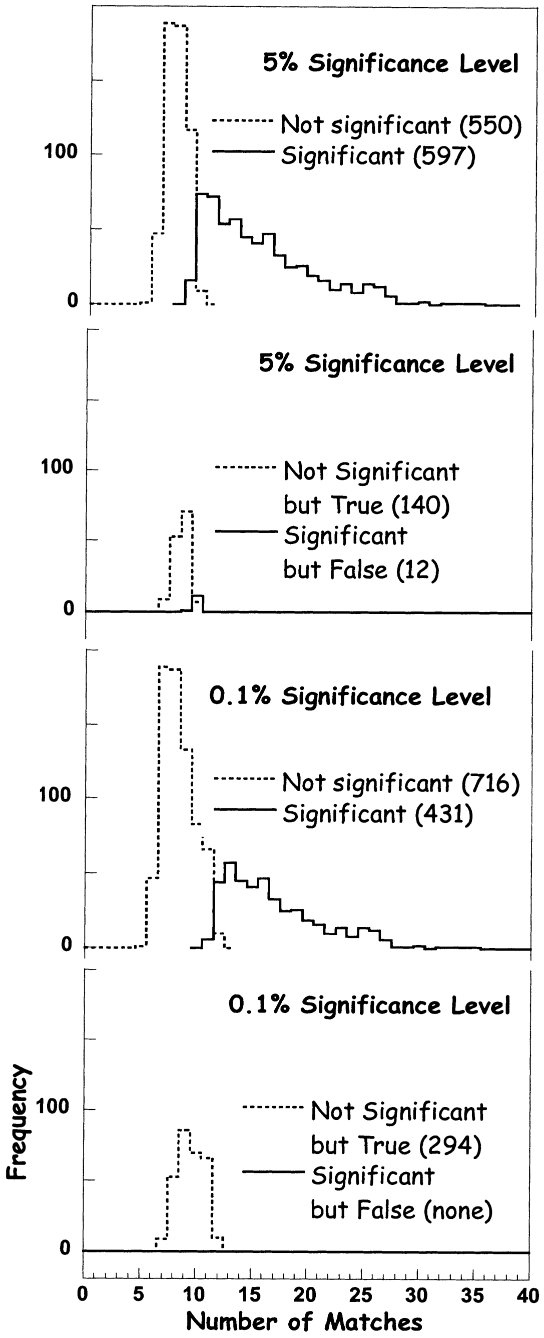

Figure 6.

4.3 Generalization and other applications of the model

The detailed and specific information on the model described in Section 2.1 refers to mass data from tryptic digests of proteins that originate from a known species that is assumed to have a correctly and completely sequenced genome. However, our model of the random matching between measured masses and theoretical

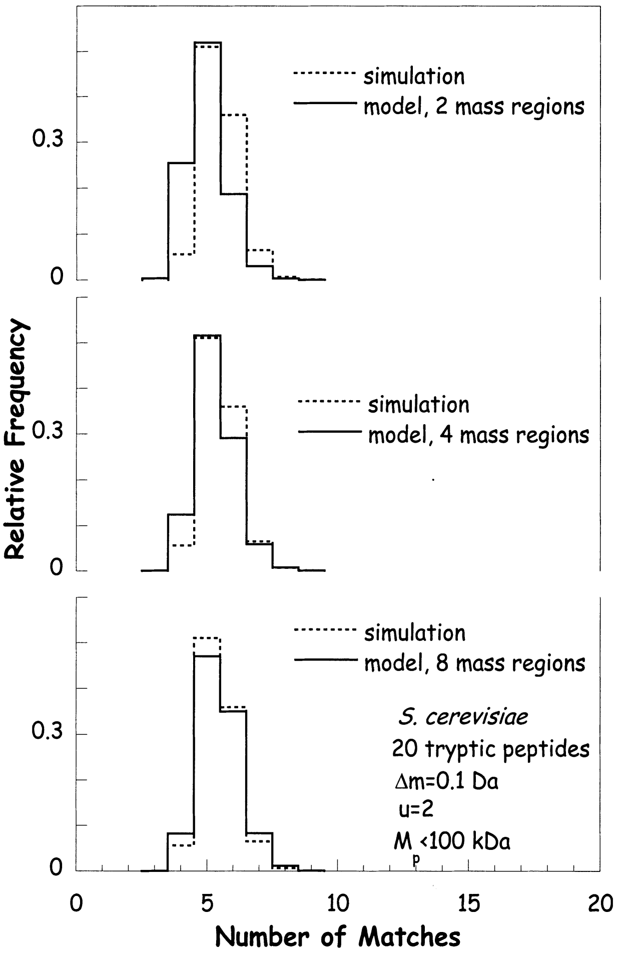

Figure 7.

masses calculated from sequence information stored in a database is general and not limited to proteins and tryptic digests, since the model only employs quantities that can be derived for any molecule with a defined sequence. The model presented here can also serve as a basis for the development of new protein identification algorithms that appropriately take the process of random matching into account in the ranking of database molecules. The strategies of the development of such algorithms will be the subject of a separate paper (Eriksson and Fenyö, in preparation).

5 Concluding remarks

We have demonstrated that rapid and accurate significance testing of mass spectrometric protein identification results can be performed by applying computations based on a simple statistical model that describes the process of random matching underlying false (random) protein identification results. The model-based approach to significance testing facilitates proteome analysis systems operating without human intervention.

This work was supported by a grant R1PM20000127001 from Proteometrics LLC, New York, NY.

Received July 24, 2001

6 References

[1] Mann, M., *Nat. Biotechnol.* 1999, *17*, 954–5.

[2] Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., *et al.*, *Science* 1995, *269*, 496–512.

[3] Goffeau, A., Barrell, B. G., Busey, H., Davis, R. W., *et al.*, *Science* 1996, *274*, 546, 563–567.

[4] Adams, M. D., Selnicker, S. E., Holt, R. A., Evans, C. A., *et al.*, *Science* 2000, *287*, 2185–2195.

[5] The Arabidopsis genome initiative, *Nature* 2001, *408*, 796–815.

[6] The *Caenorhabditis elegans* sequencing consortium, *Science* 1998, *282*, 2012–2018.

[7] Andersen, J. S., Mann, M., *FEBS Lett* 2000, *480*, 25–31.

[8] Jensen, O. N., Larsen, M. R., Roepstorff, P., *Proteins* 1998, *Suppl.* 2, 74–89.

[9] Fenyö, D., *Curr. Opin. Biotechnol.* 2000, *11*, 391–395.

[10] Jensen, O. N., Mortensen, P., Vorm, O., Mann, M., *Anal. Chem.* 1997, *69*, 1706–1714.

[11] Gras, R., Muller, M., Gasteiger, E., Gay, S., *et al.*, *Electrophoresis* 1999, *20*, 3535–3550.

[12] Henzel, W. J., Billeci, T. M., Stults, J. T., Wong, S. C., *et al.*, *Proc. Natl. Acad. Sci. USA* 1993, *90*, 5011–5015.

[13] Pappin, D. J. C., Hojrup, P., Bleasby, A., *Curr Biol* 1993, *3*, 327–332.

[14] Mann, M., Wilm, M., *Anal. Chem.* 1994, *66*, 4390–4399.

[15] Mortz, E., Vorm, O., Mann, M., Roepstorff, P., *Biol. Mass Spectrom.* 1994, *23*, 249–261.

[16] Shevchenko, A., Jensen, O. N., Podtelejnikov, A. V., Sagliocco, F., *et al.*, *Proc. Natl. Acad. Sci. USA* 1996, *93*, 14440–14445.

[17] Yates, J. R., III, *Electrophoresis* 1998, *19*, 893–900.

[18] Rabilloud, T., Kieffer, S., Procaccio, V., Louwagie, M., *et al.*, *Electrophoresis* 1998, *19*, 1006–1014.

[19] Mann, M., Hojrup, P., Roepstorff, P., *Biol. Mass Spectrom.* 1993, *22*, 338–345.

[20] Zhang, W., Chait, B. T., *Anal. Chem.* 2000, *72*, 2482–2489.

[21] Eriksson, J., Chait, B. T., Fenyö, D., *Anal. Chem.* 2000, *72*, 999–1005.

[22] Gay, S., Binz, P. A., Hochstrasser, D. F., Appel, R. D., *Electrophoresis* 1999, *20*, 3527–3534.

[23] Davies, O. L., Goldsmith, P. L. *Statistical Methods in Research and Production*, Longman Group, London 1976.

[24] Yates, J. R. D., Eng, J. K., McCormack, A. L., Schieltz, D., *Anal. Chem.* 1995, *67*, 1426–1436.

[25] Haynes, P. A., Fripp, N., Aebersold, R., *Electrophoresis* 1998, *19*, 939–945.

[26] McLafferty, F. W., Kelleher, N. L., Begley, T. P., Fridriksson, E. K., *et al.*, *Curr. Opin. Chem. Biol.* 1998, *2*, 571–578.

Appendix

Derivation of $\delta(i, \Delta m)$

The function $\delta(i, \Delta m)$ employed in Eq. (2) was derived in the following way: A mass distribution peak was chosen at 800 Da (the lowest mass considered in our model). The total number of masses in the peak was computed. A mass within the mass distribution peak was randomly chosen. The number of proteolytic peptides within $\pm\Delta m$ of that mass was counted and stored. The process beginning with randomly choosing a mass within the peak was performed 100 times. $\delta(i, \Delta m)$ was computed as the average number of matching masses divided by the total number of masses in the peak. The entire procedure was repeated for every 20th peptide mass distribution peak, up to the maximum mass considered (4500 Da). The result of this procedure is displayed in Fig. 8. The procedure was repeated for the mass accuracies: 0.2, 0.1, 0.03, 0.01, 0.003 and 0.001 Da. The mean values of $\delta(i, \Delta m)$ on each mass region (typically the four regions: 800–1045, 1046–1396, 1397–2055, 2056–4500) were computed and employed in our computations (Eq. (4)). The dependence of $\delta(i, \Delta m)$ on mass accuracy is displayed in Fig. 9, where a least-squares-fit of a function of the type shown in Eq. A1 to the computed data-points results in the parameters $a_1 = 0.0598$, $b_1 = 1.215$, $a_2 = 0.0726$, $b_2 = 1.225$, $a_3 = 0.0844$, $b_3 = 1.186$, $a_4 = 0.1159$, $b_4 = 1.207$, for the four mass regions mentioned above.

$$\delta(i, \Delta m) = 1 - \exp\left(-\left(\frac{\Delta m}{a_i}\right)^{b_i}\right) \qquad (A1)$$

Derivation of *f(S)*

Equation (4) assumes a protein-population with *H* members ranked by their number of matches with random (noncorrelated) peptide masses in a proteolytic peptide map. The resulting frequency (relative frequency) function, *f(S)*, given in Eq. (4) is identical to a probability, *P(k′)*, that at least one protein would yield exactly a score $S = k'$ random matches and that the rest of the proteins yield scores $S < k'$ random matches. Hence, Eq. (4) describes the probability that the protein or proteins characterized by the score k' would be given the highest rank and hence considered as the identification result. *P(k′)* can be expressed as:

$$P(k') = P(\text{“at least one protein yield(s) } S > (k'-1)\text{”}) - P(\text{“at least one protein yield(s) } S > (k')\text{”}) \qquad (A2)$$

The outcome complementary to “at least one protein yield(s) $S > (k' - 1)$” is: “all proteins yield $S \leq (k' - 1)$”. *Using the law of probabilities of complementary outcomes yields:*

$$\begin{aligned} P(k') &= \{1 - P(\text{“all proteins yield } S \leq (k'-1)\text{”})\} \\ &-\{1 - P(\text{“all proteins yield } S \leq (k')\text{”})\} = \\ &= \{P(\text{“all proteins yield } S \leq (k')\text{”})\} \\ &- \{P(\text{“all proteins yield } S \leq (k'-1)\text{”})\} \end{aligned} \qquad (A3)$$

Knowing the probability

$$P(\text{“a protein yields } S \leq (k')\text{”})\sum_{k=0}^{k'} pp(k),$$

and, assuming that the score of one protein is independent of that of another protein, we rewrite Eq. (A3) for *H* proteins as:

$$P(k') = f(S = k') = \left\{\sum_{k=0}^{k'} p(k)\right\}^{H} - \left\{\sum_{k=0}^{k'-1} p(k)\right\}^{H}, \text{ which}$$

is identical to Eq. (4).

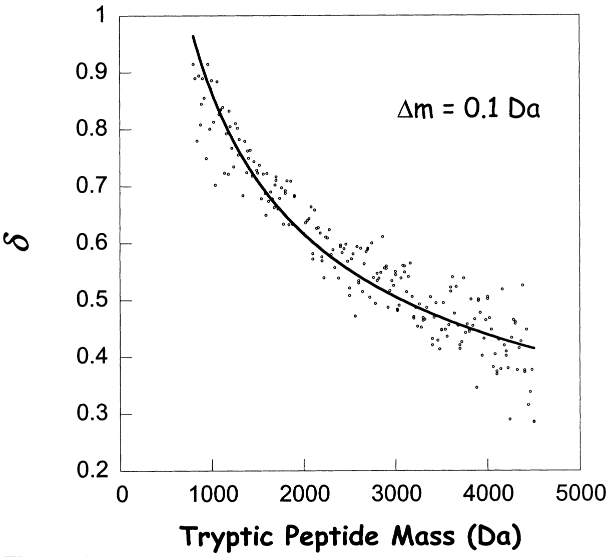

Figure 8.

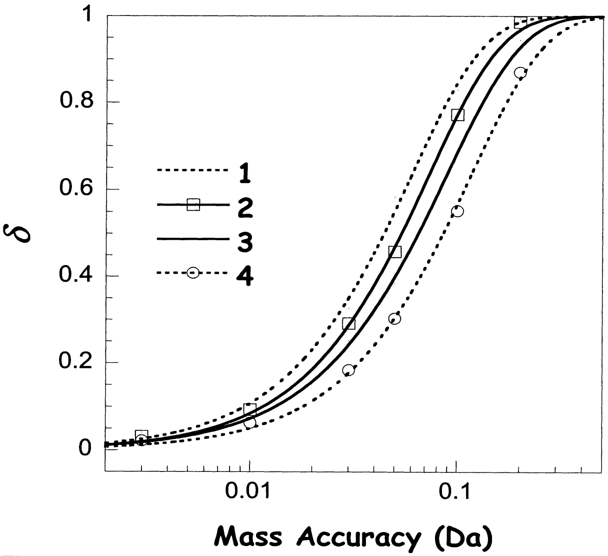

Figure 9.