A Statistical Basis for Testing the Significance of Mass Spectrometric Protein Identification Results

Jan Eriksson, Brian T. Chait, and David Fenyö*

The Rockefeller University, 1230 York Avenue, New York, New York 10021

A method for testing the significance of mass spectrometric (MS) protein identification results is presented. MS proteolytic peptide mapping and genome database searching provide a rapid, sensitive, and potentially accurate means for identifying proteins. Database search algorithms detect the matching between proteolytic peptide masses from an MS peptide map and theoretical proteolytic peptide masses of the proteins in a genome database. The number of masses that matches is used to compute a score, *S*, for each protein, and the protein that yields the best score is assumed as the identification result. There is a risk of obtaining a false result, because masses determined by MS are not unique; i.e., each mass in a peptide map can match randomly one or several proteins in a genome database. A false result is obtained when the score, *S*, due to random matching cannot be discerned from the score due to matching with a real protein in the sample. We therefore introduce the frequency function, *f*(*S*), for false (random) identification results as a basis for testing at what significance level, α, one can reject a null hypothesis, H_0: "*the result is false*". The significance is tested by comparing an experimental score, S_E, with a critical score, S_C, required for a significant result at the level α. If $S_E \geq S_C$, H_0 is rejected. *f*(*S*) and S_C were obtained by simulations utilizing random tryptic peptide maps generated from a genome database. The critical score, S_C, was studied as a function of the number of masses in the peptide map, the mass accuracy, the degree of incomplete enzymatic cleavage, the protein mass range, and the size of the genome. With S_C known for a variety of experimental constraints, significance testing can be fully automated and integrated with database searching software used for protein identification.

Protein identification by mass spectrometric (MS) peptide mapping and genome database searching has become a method of choice for the identification of proteins[1−9] from organisms with sequenced genomes.[10−13] The method is rapid, sensitive, and suitable for automation, and has been applied to a variety of tasks including the elucidation of protein function[14−18] and the determination of the composition of protein complexes.[19−21] Despite the central role that MS protein identification has assumed in proteomic research, the problem of objectively assessing the significance of identification results has remained unsolved. As increasingly complex biological problems are explored, automated identification procedures[22] become highly desirable. In such large-scale automated procedures, it becomes critical to use objective criteria for assessing the significance of each result. We report here a robust statistical solution to the problem of testing the significance of MS protein identification results.

The idea underlying MS protein identification is that a pattern of masses provides a "fingerprint" of a particular protein and that the pattern of masses can be recognized when a genome database is searched. The fingerprint can be an MS proteolytic peptide

[1] 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*, 5011−5.

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

[3] Mann, M.; Wilm, M. *Anal. Chem.* **1994**, *66* (6), 4390−9.

[4] Mortz, E.; Vorm, O.; Mann, M.; Roepstorff, P. *Biol. Mass. Spectrom.* **1994**, *23*, 249−61.

[5] Shevchenko, A.; Jensen, O. N.; Podtelejnikov, A. V.; Sagliocco, F.; Wilm, M.; Vorm, O.; Mortensen, P.; Boucherie, H.; Mann, M. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 14440−5.

[6] Jensen, O. N.; Podtelejnikov, A. V.; Mann, M. *Anal. Chem.* **1997**, *69* (9), 4741−50.

[7] Sechi, S.; Chait, B. T. *Anal. Chem.* **1998**, *70*, 5150−8.

[8] Rabilloud, T.; Kieffer, S.; Procaccio, V.; Louwagie, M.; Courchesne, P. L.; Patterson, S. D.; Martinez, P.; Garin, J.; Lunardi, J. *Electrophoresis* **1998**, *19*, 1006−14.

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

[10] *Science* **1998**, *282*, 2012−8.

[11] Blattner, F. R.; Plunkett, G., III; Bloch, C. A.; Perna, N. T.; Burland, V.; Riley, M.; Collado-Vides, J.; Glasner, J. D.; Rode, C. K.; Mayhew, G. F.; Gregor, J.; Davis, N. W.; Kirkpatrick, H. A.; Goeden, M. A.; Rose, D. J.; Mau, B.; Shao, Y. *Science* **1997**, *277*, 1453−74.

[12] Fraser, C. M.; Fleischmann, R. D. *Electrophoresis* **1997**, *18*, 1207−16.

[13] Goffeau, A.; Barrell, B. G.; Bussey, H.; Davis, R. W.; Dujon, B.; Feldmann, H.; Galibert, F.; Hoheisel, J. D.; Jacq, C.; Johnston, M.; Louis, E. J.; Mewes, H. W.; Murakami, Y.; Philippsen, P.; Tettelin, H.; Oliver, S. G. *Science* **1996**, *274*, 546, 563−7.

[14] Clauser, K. R.; Hall, S. C.; Smith, D. M.; Webb, J. W.; Andrews, L. E.; Tran, H. M.; Epstein, L. B.; Burlingame, A. L. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 5072−6.

[15] Varga-Weisz, P. D.; Wilm, M.; Bonte, E.; Dumas, K.; Mann, M.; Becker, P. B. *Nature* **1997**, *388*, 598−602.

[16] Schmidt, G.; Sehr, P.; Wilm, M.; Selzer, J.; Mann, M.; Aktories, K. *Nature* **1997**, *387*, 725−9.

[17] Yaron, A.; Hatzubai, A.; Davis, M.; Lavon, I.; Amit, S.; Manning, A. M.; Andersen, J. S.; Mann, M.; Mercurio, F.; Ben-Neriah, Y. *Nature* **1998**, *396*, 590−4.

[18] Oda, Y.; Huang, K.; Cross, F. R.; Cowburn, D.; Chait, B. T. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 6591−6.

[19] Neubauer, G.; Gottschalk, A.; Fabrizio, P.; Seraphin, B.; Luhrmann, R.; Mann, M. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 385−90.

[20] Grant, P. A.; Schieltz, D.; Pray-Grant, M. G.; Steger, D. J.; Reese, J. C.; Yates, J. R., III; Workman, J. L. *Cell* **1998**, *94*, 45−53.

[21] Wigge, P. A.; Jensen, O. N.; Holmes, S.; Soues, S.; Mann, M.; Kilmartin, J. V. *J. Cell Biol.* **1998**, *141*, 967−77.

[22] Jensen, O. N.; Mortensen, P.; Vorm, O.; Mann, M. *Anal. Chem.* **1997**, *69* (9), 1706−14.

10.1021/ac990792j CCC: $19.00 © 2000 American Chemical Society
Published on Web 02/03/2000

map—i.e., a set of masses of peptides resulting from protein digestion by an enzyme having high digestion specificity (e.g., trypsin). Identification algorithms compute the masses of peptides that individual proteins in a database would yield if they were cleaved by the same enzyme as was used in the experiment. The number of matches between masses of the experimentally obtained peptide map and the masses of the peptides from individual proteins in a database is detected. A score characterizes the result of each comparison. In some algorithms, the score is simply the number of matches,[23] whereas in other algorithms the score is the result of a computation that utilizes the number of matches as well as other criteria.[2,24] The protein or proteins yielding the best score are identified. Independent of the type of scoring system used, there is a risk of obtaining a false identification result. False results are caused by random matching of peptide masses. Each measured peptide mass can match the masses of peptides from several different proteins; i.e., an unmodified peptide will often yield random matches in addition to the match with the protein actually present in the sample, and a modified peptide will yield only random matches. A false result is obtained when the score due to random matching cannot be discerned from the score due to matching with a real protein in the sample. Hence, the distribution of the frequency of scores for protein identification by random matching must be known in order to judge the significance of protein identification results.

We derive here score frequency functions, $f(S)$, for false (random) protein identifications by *simulating* many protein identifications using random tryptic peptide maps generated from a genome database. With the null hypothesis H_0, "*the result is false*", and with $f(S)$ known, one can determine the score S_C required to reject H_0 at significance level α. To test the significance of an identification result, the experimentally obtained score, S_E, is used as the test variable. If $S_E > S_C$, H_0 is rejected. We have estimated frequency functions and the score S_C required for various significance levels for a variety of experimental constraints and for two different identification algorithms.

Significance testing can be fully automated and integrated with database searching software used for protein identification and is a general method that can be applied to any algorithm for which $f(S)$ has been determined. Hence, by using significance testing and assigning the result only by a significance level, potential confusion caused by the use of different scoring systems will be removed. The replacement of the score by the objective significance level criterion leads us to predict that significance testing will greatly facilitate automated protein identification.

MATERIALS AND METHODS

The method designed to estimate frequency functions for false protein identification involves two steps: (1) generation of 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.

Random Tryptic Peptide Maps. Random tryptic peptide maps (trypsin cleaves with high specificity at the carboxyl side of lysine and arginine residues) were generated from tryptic peptide masses predicted from the open reading frames (ORFs) of a genome database. In each map, each tryptic peptide mass was randomly chosen from a different randomly chosen protein in the database. This design of maps proves to be optimal for our purpose (see the Appendix for a background to the design). Six different sizes of maps were generated in the range of 6−80 peptides per map. A large number (=1000) of random tryptic peptide maps were generated for each peptide map size. The genome databases used were *Haemophilus influenzae*, *Saccharomyces cerevisiae*, and *Caenorhabditis elegans*, containing, respectively, 1718 (complete), 6403 (complete), and 16 332 (November 1998 release used, now complete) ORFs. The *S. cerevisiae* database was used for the majority of the present studies, while the other two databases were used to study the influence of the size of the genome on the frequency function for false protein identification.

Simulation of Protein Identification due to Random Matching. Each random tryptic peptide map was subjected to protein identification by database searching. Two different identification algorithms were employed that will be referred to as algorithm 1 and algorithm 2. Algorithm 1 ranks proteins simply by their number of matches with tryptic peptide masses in the peptide map. Algorithm 2 is a streamlined version of the "ProFound" algorithm (publicly available through the World Wide Web, http://prowl.rockefeller.edu/),[24] which ranks proteins according to a Bayesian probability calculated by comparing the measured peptide map with theoretical maps generated from the database proteins. The differences between ProFound and algorithm 2 are listed in ref 25. Algorithm 2 and ProFound take into account the number of matches between a database protein and the peptide map within the accuracy of the mass measurement, but also weigh in indirectly the protein mass as well as the assumed efficiency of the protease used in an experiment.[25] In the simulations, the score and the name of the *highest ranked* protein as well as the name of the source protein of each random tryptic peptide mass were stored for each random tryptic peptide map. This information allows "random and false" identifications to be distinguished from rare "random and true" results. If more than one protein was identified (algorithm 1 can yield two proteins with the same number of matches) and their sequences were not similar, the result was interpreted as false. A simulation with a set of different random peptide maps of the same size yields a distribution of the score for random protein identifications characteristic for that peptide map size and other constraints used in the database search. The typical parameters used in the simulations are summarized in Table 1. However, the experimentally pertinent parameters were varied systematically, one by one, to measure their respective influence on the identification score distribution.

[23] Mann, M.; Hojrup, P.; Roepstorff, P. *Biol. Mass Spectrom.* **1993**, *22*, 338−45.

[24] Zhang, W.; Chait, B. T. Submitted to *Anal. Chem.*

[25] ProFound uses Bayes' theorem: $P(k|DI) = P(k|I)\cdot P(D|kI)/P(D|I)$, where $P(k|DI)$ is the probability that the hypothesis k (protein k is the correct protein) is true given the data, D, and additional information, I. $P(k|I)$ is the prior probability, and $P(D|I)$ is a normalization factor. $P(D|kI)$ is the likelihood that the data, D, are observed for the given protein, k, and information, I. For protein identification by peptide mapping, $P(k|DI) \propto P(D|kI) \propto (\mathrm{const})^r(\Delta m^{-r})((N - r)!/N!)WF_{\mathrm{pattern}}$, where r is the number of matches, N is the number of peptides in the protein, W computes a weighing factor based on the respective difference between a measured mass and the matching theoretical mass, and F_{pattern} gives extra weight to particular peptide sequence patterns.[24] Algorithm 2 does not take into account W, F_{pattern}, and the normalization factor. The score (S) of algorithm 2 is calculated as $S = \log((\mathrm{const})^r(\Delta m^{-r})((N - r)!/N!))$.

Table 1. Typical Data and Typical Database Search Parameters Used in the Protein Identification Simulations

genome	*S. cerevisiae*
no. of tryptic peptides in a map	*n*
no. of uncleaved sites in map peptides	0
mass range in tryptic peptide maps	800−4500 Da
no. of proteins contributing to a map	*n*
no. of maps used in a simulation	≥1000
maximum no. of missed cleavage sites allowed in the database search	2
mass accuracy in database searches (Da)	0.1
maximum protein mass in data generation and database searches (kDa)	100

The code for generating peptide maps as well as for the simulation of protein identification was written in C. A script written in Perl was employed for processing the simulation results. All simulations were performed on a Dell XPS (300 MHz Pentium II) personal computer.

RESULTS

Significance Testing. A knowledge of the frequency functions, $f(S)$, for false results is the basis for testing the significance of protein identification results. Examples of $f(S)$ simulated from the *S. cerevisiae* database using random tryptic peptide maps are shown in Figure 1 (top panels). $f(S)$ is obtained if the absolute frequencies of the various scores for false results are divided by the number of random tryptic peptide maps used in the simulation.

In the simplest form of significance testing, a null hypothesis, H_0, is either rejected or not rejected at some *significance level*, α.[26] Here, we defined H_0 as "*the result is false*". The problem of testing if a protein identification result deviates significantly from H_0 falls naturally into the category of one-sided significance testing using the protein identification score S_E resulting from the experiment as the test variable. If $S_E \geq S_C$, H_0 is rejected; otherwise H_0 is not rejected. S_C will be referred to as the critical score and is derived from the relation

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

(for a discrete distribution), where α is a significance level (shaded area under $f(S)$ in Figure 1) chosen prior to the significance test.[26] α represents the statistical risk (probability) that H_0 would be rejected by the test if it actually were true. α should be small, and often the values 0.05, 0.01, and 0.001 are chosen.[26]

Critical Score. Knowledge of the critical score, S_C, is a necessary and sufficient condition for performing a significance test of a protein identification result. However, S_C must be known for the particular conditions of a given experiment. In the following paragraphs, we therefore show how S_C depends on various pertinent experimental constraints (number of mass peaks, mass accuracy, etc.) for the two algorithms used in the simulations. S_C can also be used to illustrate how the information content in an MS proteolytic peptide map depends on different experimental constraints. The information content is a measure of how easily a significant result can be achieved in a given experiment. We define the information content as $S_{\text{ideal}} - S_C$, where S_{ideal} is a score computed by assuming ideal data (n completely cleaved tryptic peptides from a single protein) and using the same constraints as for deriving S_C.

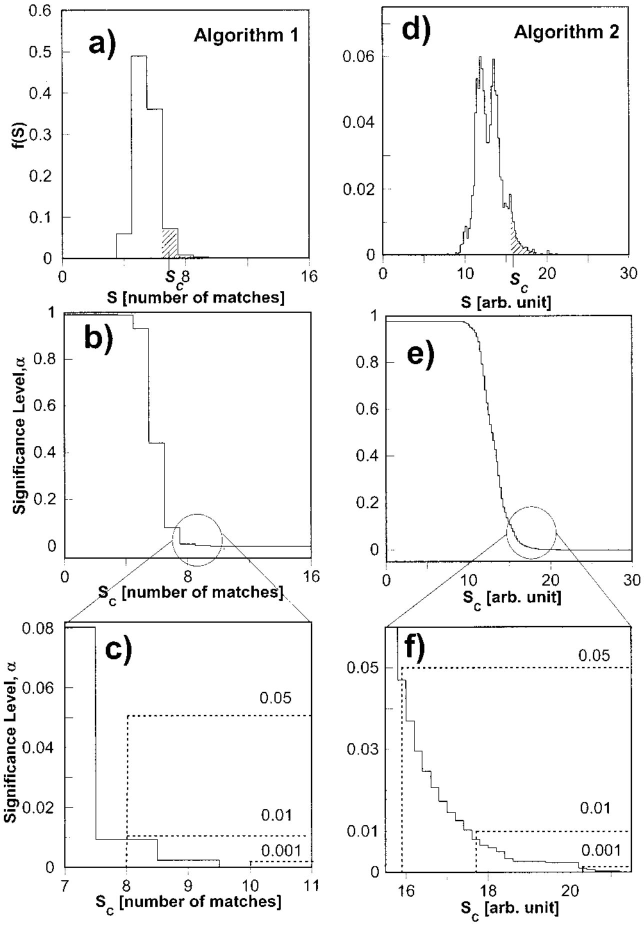

Figure 1. (a) Frequency function, $f(S)$, for false protein identification results from *S. cerevisiae* obtained by simulation with random tryptic peptide maps and algorithm 1. The area α of the shaded region under $f(S \geq S_C)$ represents the probability that a false result has a score $\geq S_C$. In significance testing, α is the significance level and is defined prior to the significance test. A protein identification result characterized by a score S is significant if $S \geq S_C$, where S_C is a critical score and corresponds, e.g., to $\alpha = 0.05$, 0.01, or 0.001. A conclusion that a result is significant ($S \geq S_C$) is equivalent to a rejection of the hypothesis H_0: "*the result is false*". The risk of rejecting H_0 if H_0 were actually true is α (the test error risk). (b) Significance level α as a function of S_C. (c) Magnified portion of (b) with horizontal lines indicating the significance levels 0.05, 0.01, and 0.001 and vertical lines indicating the corresponding critical scores. (d−f) Parallels (a−c), but with protein identification based on algorithm 2. (d) The structure in $f(S)$ reflects that the computation of the score involves the number of matches. (f) The nondiscrete nature of the score variable of algorithm 2 makes it easier to determine accurately the score S_C that corresponds to a chosen α.

Number of Masses in the Peptide Map. The number of proteolytic peptides present in a map can vary considerably between different experiments. The influence of the number of peptide mass peaks on the score S_C required for a significant result is illustrated in Figure 2. For both algorithms, S_C increases with increasing number of peptides in a map. However, the information content also increases with the size of a map. It is seen in Figure

[26] Davies, O. L.; Goldsmith, P. L. *Statistical Methods in Research and Production*; Longman Group Ltd.: London, 1976; 0-582-03040-4.

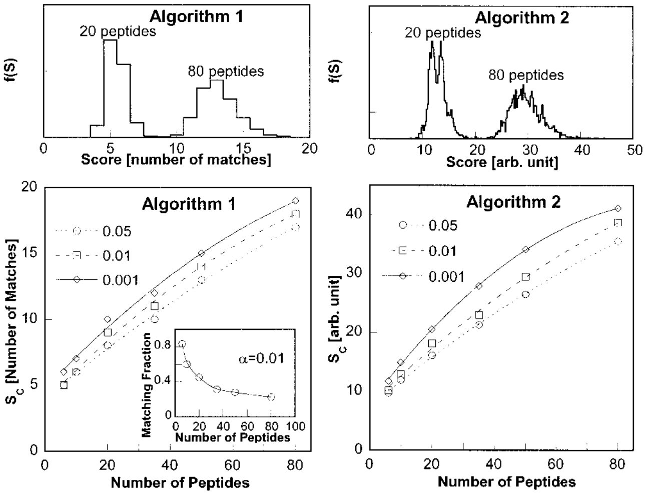

Figure 2. Top panel: $f(S)$ obtained by simulations using 3000 different random peptide maps from *S. cerevisiae* composed of 20 and 80 tryptic peptide masses, respectively. Bottom panel: Critical scores, S_C, corresponding to three different significance levels (α = 0.05, 0.01, and 0.001) as a function of the *number* of tryptic peptide masses in a peptide map. The lines through the simulated data points represent least-squares fits of second-order polynomial functions. Inset (left): Fraction of peptides in the peptide maps that match (when $\alpha = 0.01$) as a function of the number of peptides in the maps.

2 (inset, bottom left-hand panel) that the fraction of peptides in a map that match randomly decreases as a function of the number of peptides in the map. If a map is small (<6 peptides), almost all the masses must match to obtain a significant result. If the map is large, containing, e.g., 80 peptides, the result is significant ($\alpha = 0.01$) if only about 20% of the peptides match (note that these fractions change depending on the database search constraints).

Mass Accuracy. The mass accuracy, Δm, in an experiment is usually entered as a parameter in the database search. Therefore, the influence of Δm on the critical score, S_C, must be known in order to allow significance testing of any given identification result. We probed the influence of Δm on S_C by varying Δm between 0.006 and 1 Da in different simulations (state-of-the-art mass spectrometers can provide Δm < 0.1 Da for peptides). It is seen from the top panel of Figure 3 (where S_C is plotted versus Δm for algorithm 1) that the number of peptides that randomly match the identified protein decreases sharply with decreasing Δm for $0.8 < \Delta m < 1$ and for $\Delta m \leq 0.1$ Da. The observed Δm dependence of the random matching of peptide masses can be understood from the fact that peptides are composed of only a few different types of atoms that when combined always yield peptide masses that cluster in mass regions ~0.25 Da wide, with a ~0.75 Da mass region between the clusters devoid of peptide masses.[27,28] The decrease of S_C around 1 Da is due to discrimination against adjacent mass clusters, while the decrease below 0.1 Da arises because a decreasing fraction of the masses within a cluster can match. For algorithm 1, the score corresponding to ideal data, S_{ideal}, is independent of Δm (Figure 3, top; all n peptides in the map match). For algorithm 2, S_{ideal} and S_C depend on Δm (Figure 3, bottom). $S_{\mathrm{ideal}} \propto n \log(\Delta m^{-1})$[24] and $S_C \propto r_{\mathrm{random}} \log(\Delta m^{-1})$, where r_{random} is the number of random matches. Since $r_{\mathrm{random}} < n$, the information content ($S_{\mathrm{ideal}} - S_C$) always increases with decreasing Δm. Hence, for both algorithms a reduction of Δm facilitates significant protein identification.

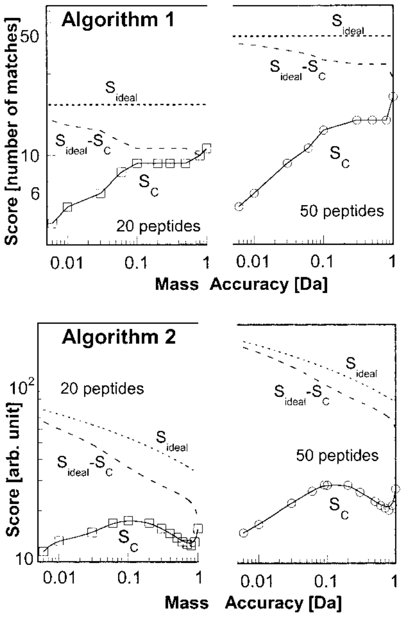

Figure 3. Critical score, S_C, corresponding to $\alpha = 0.01$ as a function of the *mass accuracy* for tryptic peptide maps from *S. cerevisiae* with 20 and 50 peptide masses. Scores corresponding to ideal data, S_{ideal} (see the text), are shown for comparison. Also shown is a plot of $S_{\mathrm{ideal}} - S_C$, demonstrating that higher accuracy (i.e., lower Δm) facilitates significant protein identification. Top: Algorithm 1. Bottom: Algorithm 2. Scores for ideal data and algorithm 2 were computed for a 45 kDa protein.

Incomplete Enzymatic Cleavage. Enzymatic digestion is often incomplete. Therefore, the expected highest number u of specific sites not cleaved in a peptide is typically entered as a constraint in the database search. The influence of u on the critical score, S_C, must therefore be established to allow significance testing of any given identification result. In the present work, u was varied between 0 and 4 in different simulations. It is seen from the results in Figure 4 (top panel), where S_C for algorithm 1 is plotted as a function of u, that the more complete the cleavage the lower the number of random matches characterizing the false results. S_C for algorithm 2 (Figure 4, bottom panel) saturates with increasing u due to an intrinsic moderation of the computed score, S, by the number, N, of possible proteolytic peptide masses in each individual database protein. $N \propto 1 + u$ and $S \propto \log((N - r)!/N!) \approx \log(N^{-r})$, where r is the number of matches.[24] Assuming n matches for ideal data (n completely cleaved peptides from a single protein) and r_{random} matches for random data yield $S_{\mathrm{ideal}} - S_C \propto (n - r_{\mathrm{random}})(\mathrm{const} + \log((1 + u)^{-1})$, which decreases slowly with increasing u (Figure 4, bottom panel).

For both algorithms, it is observed that the change of S_C is most pronounced at low values of u and that complete cleavage facilitates significant protein identification. However, we note that the use of specific mass patterns that sometimes arise from incomplete cleavage can yield additional information that proves to be highly constraining for protein identification by MS proteolytic peptide mapping.[24,29]

[27] Mann, M. *43rd ASMS Conference on Mass Spectrometry and Allied Topics*; Atlanta, Georgia, May 21–26, 1995.

[28] Fenyo, D.; Qin, J.; Chait, B. T. *Electrophoresis* **1998**, *19*, 998–1005.

[29] Jensen, O. N.; Vorm, O.; Mann, M. *Electrophoresis* **1996**, *17*, 938–44.

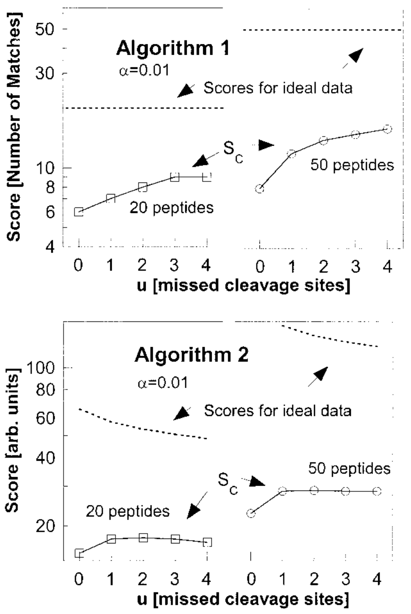

Figure 4. Critical score, S_C ($\alpha = 0.01$), for tryptic peptide maps with 20 and 50 masses versus the maximum number of missed cleavage sites, *u*, that was allowed in the *S. cerevisiae* database search. Scores corresponding to ideal data, S_{ideal} (see the text), are shown for comparison. Top: Algorithm 1. Bottom: Algorithm 2. Scores for ideal data and algorithm 2 were computed for a 45 kDa protein.

Protein Mass. The protein mass can be used as a constraint in the database search. This constraint is usually obtained from SDS−gel electrophoresis and should be used with caution, since protein degradation and anomalous migration of modified proteins can yield misleading molecular weights. For simplicity, most of our simulations were restricted to protein masses of 100 kDa (for generating the peptide maps as well as for the database search). About 95% of *S. cerevisiae* proteins are within this mass range. To cover the remaining 5% of the proteins, we studied the influence on the critical score, S_C, as a function of the protein mass range. Figure 5 shows that algorithm 1 yields a high degree of random matching with high-mass proteins, whereas S_C for algorithm 2 is less sensitive to an increased protein mass range. A larger protein mass implies a larger number, *N*, of possible proteolytic peptides. As discussed above, algorithm 2 moderates the score with increasing *N*. Algorithm 1 lacks this feature and therefore favors false identification of large proteins.

Genome Size. The critical score, S_C, was studied as a function of the size of the genome. The results shown in Figure 6 are based on data (random tryptic peptide maps) from a prokaryote, *H. influenzae*, a single-cell eukaryote, *S. cerevisiae* (budding yeast), and a multicellular organism, *C. elegans* (nematode). It is seen from Figure 6 that S_C increases with the size of the genome. The increase of S_C is steeper the lower the size of the genome. The small difference between the respective S_C values for *S. cerevisiae* and *C. elegans* implies that protein identification by peptide mapping can also be accomplished for larger genomes without necessarily reaching problematic frequencies of false results. A reduced database of *C. elegans* proteins was generated by randomly selecting 6.4×10^3 ORFs (the same number as in the entire *S. cerevisiae* genome) from the *C. elegans* genome. Simulations with the reduced *C. elegans* database yielded a value of S_C indistinguishable from that of *S. cerevisiae* (data not shown). This finding implies that the saturation of S_C as a function of genome size is not due to any special feature of *C. elegans* or *S. cerevisiae* proteins. The score frequency function for all results (false and true) was essentially independent of whether maps generated from the reduced *C. elegans* database were used to search the *S. cerevisiae* database or the reduced *C. elegans* database. This result is due to the highly similar distribution of tryptic peptide masses for different genomes.[28] Therefore, the dependence of S_C on the genome size shown here can be used to estimate S_C for any genome within the size range studied.

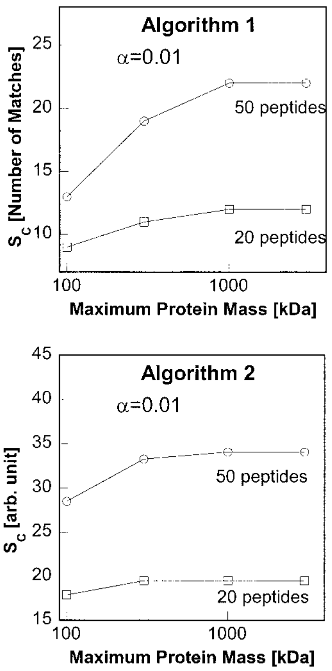

Figure 5. Critical score, S_C, that yields $\alpha = 0.01$ for tryptic peptide maps with 20 and 50 masses as a function of the *maximum protein mass* that was allowed in the *S. cerevisiae* database search. Top: Algorithm 1. Bottom: Algorithm 2.

DISCUSSION

Statistical Uncertainties. The number of protein identifications simulated and the shape of the score frequency function can influence the accuracy of S_C, the score required for a significant result. We probed how S_C varied due to statistical fluctuations by (1) repeated simulation with the same number of maps using identical conditions except for the set of random numbers employed to generate the maps and (2) by varying the number of random peptide maps used per simulation. The pronounced discrete nature of the frequency function of algorithm 1 implies an inherent sensitivity to statistical fluctuations in the simulations. If S_C is statistically well determined, it should converge to a particular value as the number of maps in the simulation is increased. This was not observed for algorithm 1 for $\alpha = 0.01$ or $\alpha = 0.001$. Instead, S_C fluctuated by one match as the number of maps used was increased in six steps from 250 to 15 000. In these instances, the highest S_C observed was assumed as the result. The relative importance of this uncertainty decreases with increasing size of the peptide maps, because of the larger number of random matches (Figure 2).

The approach of fitting analytical functions to frequency functions could be a means of reducing difficulties associated with

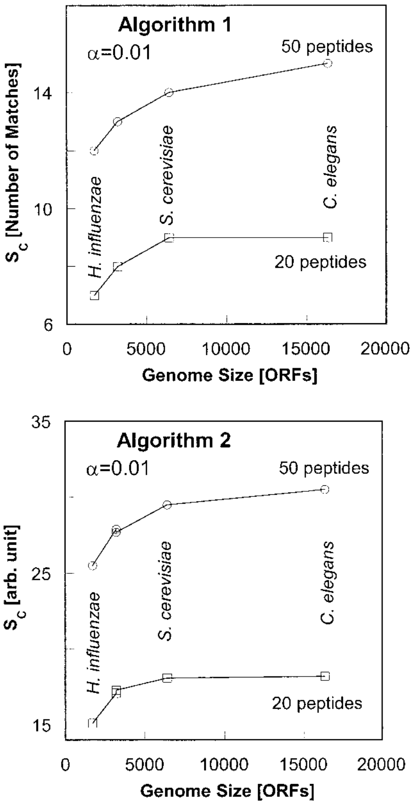

Figure 6. Influence of the *size of the genome database* on the critical score, S_C, required for $\alpha = 0.01$ and tryptic peptide maps with 20 and 50 masses. The simulated data were obtained by using the *H. influenzae*, *S. cerevisiae* (yeast), and *C. elegans* databases. The data point between *H. influenzae* and yeast was obtained by randomly dividing the yeast genome into two parts of equal size. Top: Algorithm 1. Bottom: Algorithm 2.

discrete distributions and statistical fluctuations. Although this approach has been employed for determination of scores required for significance in sequence or structure comparison algorithms,[30] it remains to be explored for simulation of random protein identification.

The nondiscrete nature of algorithm 2 allows even minor statistical fluctuations to be resolved and examined by performing multiple simulations under identical conditions using peptide maps generated from different series of random numbers. Thus, the fluctuation of S_C due to different responses to different random data could be probed. The standard deviation of the mean S_C derived from five different simulations decreased sharply when the number of maps used per simulation was increased from 500 to 1000, and then changed very slowly when the number of maps was further increased to 15 000. For five simulations each using 1000 maps with 20 random tryptic peptide masses, the relative standard deviation from the mean S_C was 0.6%, 1.5%, and 2.5% for the 0.05, 0.01, and 0.001 significance levels, respectively. Hence, for algorithm 2 the magnitude of S_C appears to be well established at 1000 maps per simulation.

Exploring a Large Parameter Space. The results presented in Figures 1−6 are based on a large number of simulations involving $>10^5$ protein identifications. Although these simulations represent only a small fraction of the range of search parameters studied, we found that estimation based on the derived functions that describe how S_C varies in the parameter space (Figures 2−6) provides an accurate procedure to assess S_C in an arbitrary point in the parameter space. We tested this approach by comparing such estimations of S_C with values of S_C derived from direct simulations for randomly chosen points in the parameter space using algorithm 2. The deviations between the estimated and the simulated values of S_C were within the observed standard deviation of S_C discussed above.

Use of Significance Testing. We will discuss briefly what significance testing can do as well as what it cannot do when applied to protein identification. In contrast with the identification score, the significance level of a protein identification result gives an objective view of the quality of the result. However, it should be noted that significance testing can *never* definitively prove whether a result is true or false. A significant result is either false or true, as is a nonsignificant result. The significance level is the calculated risk of obtaining a false result in a single identification. The relative frequency of false results for a group of identifications depends on the data as well as on the significance level chosen. If α is decreased, the relative frequency of false results is expected to decrease. However, choosing a very low α can sometimes lead to an increase of the relative frequency of true results considered nonsignificant. Optimized protein identification requires (1) the use of an identification algorithm that maximizes the relative frequency of true identifications and (2) the use of significance testing at an appropriate significance level to discriminate against false identifications. We will discuss the details of such optimization in a separate paper.[31] Here, we simply emphasize that significance testing has the potential to reduce the relative frequency of false identifications independent of the identification algorithm used.

If significant protein identification is not achieved directly by the described peptide mapping procedure, a researcher can try to obtain further additional experimental information that provides additional identification constraints. A good source of such information is tandem mass spectrometry,[32−34] which utilizes fragmentation of given proteolytic peptide ions in the mass spectrometer followed by analysis of the resulting fragment ion masses and database searching. Results obtained from tandem MS (and other experimental constraints used) should also be subjected to significance testing once a statistical basis has been established by simulation.

CONCLUSIONS

We have shown that computer simulations of random protein identifications can provide a statistical basis for testing the significance of protein identification results. The frequency function for false results derived from the simulations can be employed to find the score S_C required to reject a hypothesis of false protein identification at some significance level. We have investigated how S_C varies with various pertinent experimental constraints, and have established that these functions can be used to estimate the value

[30] Levitt, M.; Gerstein, M. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 5913−20.

[31] Eriksson, J.; Chait, B. T.; Fenyo, D. Manuscript in preparation.

[32] Yates, J. R., III; Eng, J. K.; McCormack, A. L.; Schieltz, D. *Anal. Chem.* **1995**, *67* (7), 1426−36.

[33] Haynes, P. A.; Fripp, N.; Aebersold, R. *Electrophoresis* **1998**, *19*, 939−45.

[34] McLafferty, F. W.; Kelleher, N. L.; Begley, T. P.; Fridriksson, E. K.; Zubarev, R. A.; Horn, D. M. *Curr. Opin. Chem. Biol.* **1998**, *2*, 571−8.

of S_C. Hence, the statistical framework presented here can be integrated with protein identification algorithms and fully automated. We envision that, in the future, protein identification results will be characterized by a significance level rather than by a score.

ACKNOWLEDGMENT

Financial support to J.E. was provided by the Wenner-Gren Center Foundation. This work was funded in part by a grant from the National Institutes of Health (RR00862). We thank M. Cadene, J. C. Padovan, M. Trester, R. Wang, and W. Zhang for fruitful discussions.

APPENDIX

Design of Tryptic Peptide Maps. The computer-generated data used in our protein identification simulations were random tryptic peptide maps, i.e., maps where each tryptic peptide mass was randomly generated from a different randomly selected protein. These maps were used with the specific goal of elucidating the score frequency function, $f(S)$, for false (random) identification results.

A completely different but formally correct alternative way of studying the scores of false results is to construct *ideal* proteolytic peptide maps, each with all masses from a *single* randomly selected protein, perform simulations, and use the score frequencies of the *second highest ranked* protein resulting from the database search as an estimate of $f(S)$. In Figure 7, $f(S)$ derived from the second highest ranked proteins when using ideal maps with 20 tryptic peptides is compared with $f(S)$ derived from the scores of the *highest ranked* proteins identified on the basis of *random* tryptic peptide maps with 20 tryptic peptides. It is seen that the two approaches yield very similar results. However, if the number of peptide masses of the ideal maps is large, only high-mass proteins can contribute to the maps (maximum number of tryptic peptides ≈ protein mass [Da]/1500 [Da]). High-mass proteins have low abundance in a genome, and therefore, the number of *different* large ideal maps is limited. This limitation would obscure the statistical quality of the score distribution for random matching. In contrast, random tryptic peptide maps can be generated from the entire database. We therefore chose the approach with random tryptic peptide maps in our simulations.

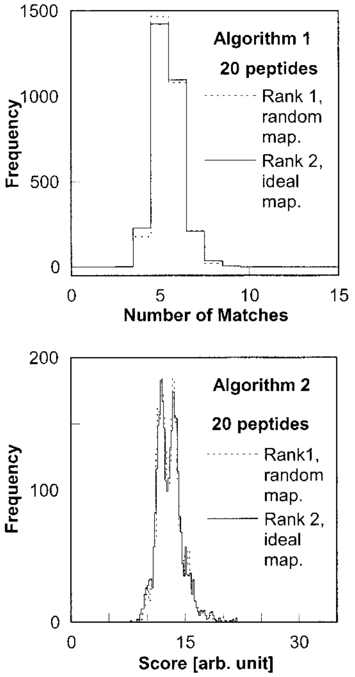

Figure 7. Score frequencies due to random matching from two different simulation models using algorithm 1. Similar score frequencies are obtained for the *highest ranked* protein when using *random* tryptic peptide maps (each tryptic peptide mass from a different protein) as are obtained for the *second highest ranked* protein when using different *ideal* tryptic peptide maps (all tryptic peptide masses from a single randomly selected protein). We note that the latter method is not practical for use with large peptide maps (see the text for details).

Received for review July 19, 1999. Accepted November 22, 1999.

AC990792J