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

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

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, \( \alpha \), 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 \( \alpha \). 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 \(^3\)–\(^9\) 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

---

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, whereas in other algorithms the score is the result of a computation that utilizes the number of matches as well as other criteria. 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 trypptic 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 $\alpha$. To test the significance of an identification result, the experimentally obtained score, $S_\text{obs}$, is used as the test variable. If $S_C > S_\text{obs}$, $H_0$ 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 trypptic peptide maps (trypsin cleaves with high specificity at the carboxyl side of lysine and arginine residues) were generated from trypptic peptide masses predicted from the open reading frames (ORFs) of a genome database. In each map, each trypptic 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 trypptic 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 16332 (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 trypptic 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 trypptic 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/), 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 weight in indirectly the protein mass as well as the assumed efficiency of the protease used in an experiment. 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 trypptic peptide map were stored for each random trypptic 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.

(25) ProFound uses Bayes’ theorem: $P(k|D)=P(D|k)P(k)/P(D|k)P(D)$, where $P(k|D)$ is the probability that the hypothesis $k$ (protein $k$ is the correct protein) is true given the data, $D$, and additional information. $P(k)$ is the prior probability, and $P(D|k)$ is a normalization factor. $P(D|k)$ is the likelihood that the data, $D$, are observed for the given protein, $k$, and information, $I$. For protein identification by peptide mapping, $P(D|k)=\phi_I(N-1)\phi_W\phi_{\text{match}}$, where $m$ 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 $\phi_{\text{match}}$ gives extra weight to particular peptide sequence patterns. Algorithm 2 does not take into account $W$, $\phi_{\text{match}}$, and the normalization factor. The score $S$ of algorithm 2 is calculated as $S = \log(\phi_I(N-1)\phi_W\phi_{\text{match}})/N$.
Table 1. Typical Data and Typical Database Search Parameters Used in the Protein Identification Simulations

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>genome</td>
<td>S. cerevisiae</td>
</tr>
<tr>
<td>no. of tryptic peptides in a map</td>
<td>n</td>
</tr>
<tr>
<td>no. of uncleaved sites in map peptides</td>
<td>0</td>
</tr>
<tr>
<td>mass range in tryptic peptide maps</td>
<td>800–4500 Da</td>
</tr>
<tr>
<td>no. of proteins contributing to a map</td>
<td>n</td>
</tr>
<tr>
<td>no. of maps used in a simulation</td>
<td>≥1000</td>
</tr>
<tr>
<td>maximum no. of missed cleavage sites allowed in the database search</td>
<td>2</td>
</tr>
<tr>
<td>mass accuracy in database searches (Da)</td>
<td>0.1</td>
</tr>
<tr>
<td>maximum protein mass in data generation and database searches (kDa)</td>
<td>100</td>
</tr>
</tbody>
</table>

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 M Hz 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, \( \alpha \). 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

\[
\sum_{S \leq S_C} f(S) \leq \alpha
\]

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

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 \( S_{\text{ideal}} \) 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 \( S_C \).

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

---

Jensen, O. N.; Vorm, O.; Mann, M.

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, $\Delta m$, in an experiment is usually entered as a parameter in the database search. Therefore, the influence of $\Delta m$ 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 $\Delta m$ on $S_C$ by varying $\Delta m$ between 0.006 and 1 Da in different simulations (state-of-the-art mass spectrometers can provide $\Delta m < 0.1$ Da for peptides).

2 (top panel), where $S_C$ is plotted versus $\Delta m$ for algorithm 1) that the number of peptides that randomly match the identified protein decreases sharply with decreasing $\Delta m$ for $0.8 < \Delta m < 1$ and for $\Delta m \leq 0.1$ Da. The observed $\Delta m$ 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 $\sim 0.25$ Da wide, with a $\sim 0.75$ 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 $\Delta m$ (Figure 3, top; all $n$ peptides in the map match). For algorithm 2, $S_{ideal}$ and $S_C$ depend on $\Delta m$ (Figure 3, bottom). $S_{ideal} \propto n \log(\Delta m^{-1})$24 and $S_C \propto r_{random} \log(\Delta m^{-1})$, where $r_{random}$ is the number of random matches. Since $r_{random} < n$, the information content ($S_{ideal} - S_C$) always increases with decreasing $\Delta m$. Hence, for both algorithms a reduction of $\Delta m$ facilitates significant protein identification.

**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.

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

---


(29) Jensen, O. N.; Vorm, O.; Mann, M. Electrophoresis 1996, 17, 938–44.
**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 $6.4 \times 10^5$ 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.

**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
described 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 $\alpha$ is decreased, the relative frequency of false results is expected to decrease. However, choosing a very low $\alpha$ 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. 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, 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

---


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 \( \approx \) 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.

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

AC990792J