research articles Journal of proteome research

The Statistical Significance of Protein Identification Results as a Function of the Number of Protein Sequences Searched

Jan Eriksson*,† and David Fenyö‡,§

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

Received March 24, 2004

The potential for obtaining a true mass spectrometric protein identification result depends on the choice of algorithm as well as on experimental factors that influence the information content in the mass spectrometric data. Current methods can never prove definitively that a result is true, but an appropriate choice of algorithm can provide a measure of the statistical risk that a result is false, i.e., the statistical significance. We recently demonstrated an algorithm, Probity, which assigns the statistical significance to each result. For any choice of algorithm, the difficulty of obtaining statistically significant results depends on the number of protein sequences in the sequence collection searched. By simulations of random protein identifications and using the Probity algorithm, we here demonstrate explicitly how the statistical significance depends on the number of sequences searched. We also provide an example on how the practitioner's choice of taxonomic constraints influences the statistical significance.

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

Proteome analysis projects emerge in the wake of completed genome projects.[1−4] Proteome analysis reveals what genes are expressed in the cells, and more detailed experiments can provide information on differential expression, subcellular localization, protein interaction,[5−10] and post-translational modifications.[11−15] The state-of-the art approach for the identification of the genes expressed (so-called protein identification) utilizes mass spectrometry (MS) of proteolytically digested proteins in combination with the searching of a collection of protein sequences derived from genomic information. The plentitude of genomes sequenced allows laboratories to analyze many different proteomes. We here demonstrate how the difficulty of obtaining a statistically significant identification result depends on the number of protein sequences searched, and hence, we also elucidate that proteome analysis procedures should be optimized for each respective organism studied.

Mass spectrometric protein identification relies on the procedure of comparing experimental mass data, acquired from a sample containing proteolytic peptides, with theoretical mass information obtained from a collection of protein sequences.[16−20] Different algorithms are available for the process of finding the protein sequence that yields the best match with experimental mass data.[21,22] The potential for obtaining a true identification result depends on the choice of algorithm as well as on experimental factors that influence the information content in the mass spectrometric data. Accurate information on the statistical significance, i.e., the risk that a result is false, is critical to any serious proteome analysis effort.[23] We recently demonstrated an algorithm, Probity, which assigns the statistical significance to each result.[24] We showed by simulations, employing random data and searching of the *S. cerevisiae* protein sequence collection, that Probity does not favor particular protein sequences. We also demonstrated that the distribution of scores assigned to the highest ranked proteins for a set of random mass data is independent of what experimentally related constraints are employed in the searching (mass accuracy, number of missed cleavage sites). Although experimental factors as well as any features of a sequence collection can be appropriately accounted for in an algorithm, the risk of obtaining a good match of a protein sequence by chance with an experimental data set must always depend on the number of sequences searched. The more sequences searched the higher the risk of a good match by chance.

Knowing that Probity accounts for experimentally related constraints, we here employed this algorithm to simulate the explicit influence of the number of protein sequences searched on the statistical significance of identification results. The results also elucidate the importance of keeping searching of a sequence collection focused to precisely the organism of interest in each respective experiment.

Materials and Methods

The *H. influenzae*, *S. cerevisiae*, *C. elegans*, and *H. sapiens* proteomes containing 1.7×10^3, 6.4×10^3, 1.9×10^4, and 5.3×10^4 sequences, respectively, were employed for simulating protein identification. Protein digestion was performed in silico

* To whom correspondence should be addressed. E-mail: jan.eriksson@kemi.slu.se.
† Swedish University of Agricultural Sciences.
‡ Amersham Biosciences.
§ The Rockefeller University.

10.1021/pr0499343 CCC: $27.50

Published on Web 08/27/2004

assuming exposure to trypsin (cleaves with high specificity at the carboxyl side of lysine and arginine residues). Peptides within a mass region between 800 and 4500 Da were considered. Sets of randomly selected proteolytic peptide masses were generated as described in refs 23−25. Each random data set contained 20 peptide masses. A total of 900−2000 different random protein identifications were performed employing the Probity algorithm for each respective sequence collection (organism) and using the following search constraints: $\Delta m = 0.03$ Da (mass accuracy), $u = 1$ (maximum number of missed cleavage sites), and $M < 100$ kDa (protein mass).

Scripts written in *Perl* and programs written in *C* were employed for the simulations, which were performed on a personal computer (Dell Pentium IV, 2.66 GHz).

Algorithm. Previously, we presented accurate descriptions of the ranking method in the Probity algorithm.[24] We demonstrated that Probity responds to random data randomly and that the score distribution for random data is independent of the search constraints Δm and u. We also showed that the statistical significance of the score of an individual identification result could be computed by taking into account features of the collection of sequences searched. In this paper, we simplify the formula for computing the significance of a result, and thereby we can also demonstrate in a more straightforward way how the significance depends on the number of sequences searched. The probability that an individual protein sequence having k_u proteolytic peptides will yield *at least* k' matches by chance is

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

where *p(k)* is given by eq 2

$$p(k) = \sum_{k_i, \sum k_i = k} \left\{ \prod_{i=1}^{q} \binom{n_i}{k_i} \cdot p_i'^{k_i} \cdot (1 - p_i')^{n_i - k_i} \right\} \tag{2}$$

where the index *i* denotes a proteolytic peptide mass region. Eight different mass regions were used here ($q = 8$ in eq 2) to cover the mass range (800−4500 Da). Proteolytic peptide masses are distributed in narrow regions around each nominal mass value. The widths and heights of these distributions vary with mass.[25] In Probity, it is assumed that within each proteolytic peptide mass region the heights and widths of the mass distribution peaks are constant. The choice of the number of mass regions, *q*, influences the accuracy of the computations,[25] and here we use $q = 8$. The symbol n_i denotes the number of experimental proteolytic peptide masses experimentally observed in mass region *i*. The probabilities p'_i of eq 2 are given by

$$p'_{ik_i} = f_i \frac{k_u}{m_{i+1} - m_i} \delta(i, \Delta m, u) \tag{3}$$

where $m_{i+1} - m_i$ denotes the number of mass distribution peaks in region *i* and f_i denotes the fraction of the total number of peptides that are estimated to belong to region *i* in the sequence collection. The fraction f_i of the proteolytic peptides from an individual protein that on average falls into *i* is estimated from the fraction of the total number of proteins in the sequence collection yielding peptides within *i*. $\delta(i, \Delta m, u)$ denotes a function that depends on the shape of the peptide mass distribution peak. $\delta(i, \Delta m, u)$ can be determined by simulation.[25]

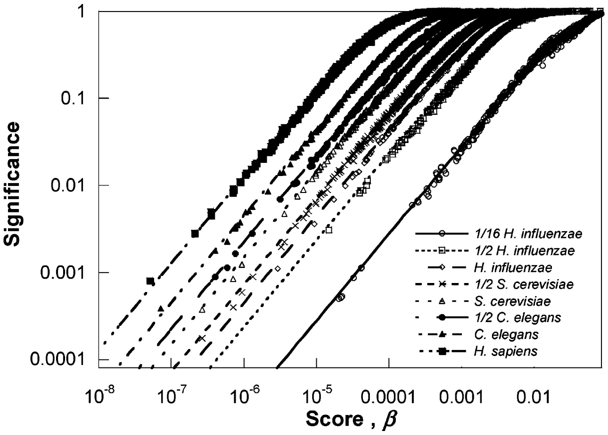

Figure 1. Statistical significance displayed as a function of the score, β, of the highest ranked protein sequence for the Probity algorithm when searching various protein sequence collections. The data points result from the use of eq 4, and the lines represent least-squares fits to these data of the functions of eq 6, $1 - (1 - a\beta)^N$, where N is the number of sequences in each respective collection searched and a is the fitting parameter.

The statistical significance, *S*, can be expressed as a function of β[24]

$$S = P(\text{at least 1 protein will yield} \le \beta) = 1 - P(\text{all proteins yield} > \beta) = 1 - \prod_{k_u=1}^{k_u^{\max}} P_{k_u}(>\beta) \tag{4}$$

with

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

where $k_{\max}$ is determined by β, and ψ_{ku} is the frequency of a particular k_u value in the collection of sequences used. The probability, P_{ku}, in eq 5 is determined by β in such a way that the summation is continued until $1 - P_{ku}$ becomes smaller than β. If P_{ku} were continuous functions they would assume the value of $1 - \beta$ for all β. Since P_{ku} (and *p(k)*) are discrete functions there is always a deviation between P_{ku} and $1 - \beta$, such that $P_{ku} > 1 - \beta$. We here express this as

$$P_{k_u} = 1 - (1 - a_{k_u} \cdot \beta)^{\Psi_{k_u}}$$

If we assume that $a_{k_u} = a =$ constant and express the sum of all Ψ_{k_u} as N, we can express the significance as

$$S = 1 - (1 - a\beta)^N \tag{6}$$

where *N* means the total number of sequences searched.

Results

The relation between the statistical significance and the number of sequences searched was obtained by simulating many protein identifications using random data sets, the Probity algorithm (eqs 1−5), and eight different sizes of the sequence collections searched (four sequence collections were obtained by randomly selecting sequences of *H. influenzae, S. cerevisiae*, and *C. elegans*, referred to as: 1/16 *H.i*, $^1/_2$ *H.i.*, $^1/_2$ *S.c*, and $^1/_2$ *C.e*, where 1/16 and $^1/_2$ denote the fractions selected of the respective total number of sequences). Figure 1 displays the simulation results together with least-squares fits to the functions expressed in eq 6 with a as the fitting parameter. The resulting values of the parameter a were similar in all cases and yielded a mean value of a of 0.270 with a standard deviation of 0.011.

The relationship established between β and S has two important consequences: (1) Equation 6 can be used instead of eq 5, which improves the computational speed of the Probity algorithm. (2) The difficulty of obtaining a significant result can be estimated for any collection of sequences for which N is known.

The consequence (2) can be viewed as a predictive power. Figure 2 displays the predicted statistical significance of equally well matching results when searching in sequence collections with different numbers of sequences. β' values that yielded the respective significance values $S' = 0.001$ and $S' = 0.01$ in the *S. cerevisiae* sequence collection were entered in eq 6 together with the N values of the respective sequence collection. It is seen Figure 2. that e.g data with a degree of matching that yield a 0.1% significance level in *S. cerevisiae* would yield only about 1% significance in *H. sapiens.*

In Figure 3 we show the influence of the choice of taxonomic precision when selecting the sequence collection to search. The results displayed in Figure 3 were computed in the following way: First, the β-value, β', of *S. cerevisiae* data that yields 0.1% significance when searching the *S. cerevisiae* sequence collection only was derived. The N-values of sequence collections containing all current fully sequenced fungi genomes (4) and all fully sequenced eukaryota genomes (19) respectively were estimated from the number of ORFs displayed in the GOLD database (genomes on line, http://wit.integratedgenomics.com/GOLD/). β' and the respective N values were entered into eq 6 and the significance values for the respective sequence collection was computed. It is seen in Figure 3 that changing from *S. cerevisiae* and a significance level of 0.1% to all fungi results in a 1% significance level and that changing to all eukaryota yields a 5% significance level.

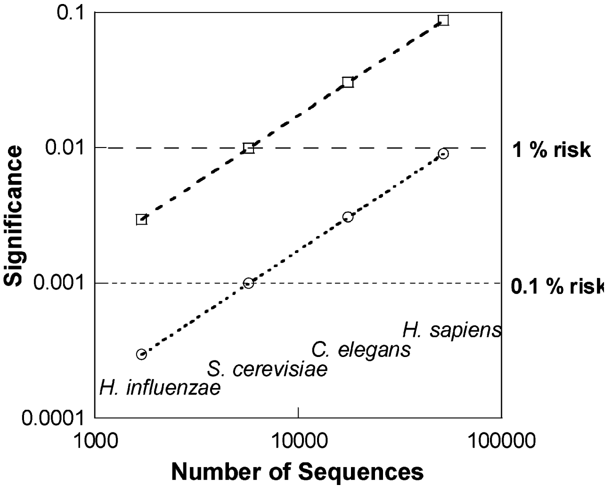

Figure 2. Comparison of the statistical significance as obtained from eq 6 for data of the same quality (same score) for different organisms. The results for *S. cerevisiae* are displayed as reference points at the 0.1% significance level (lower part) and at the 1% significance level (upper part). The lines are guides for the eye.

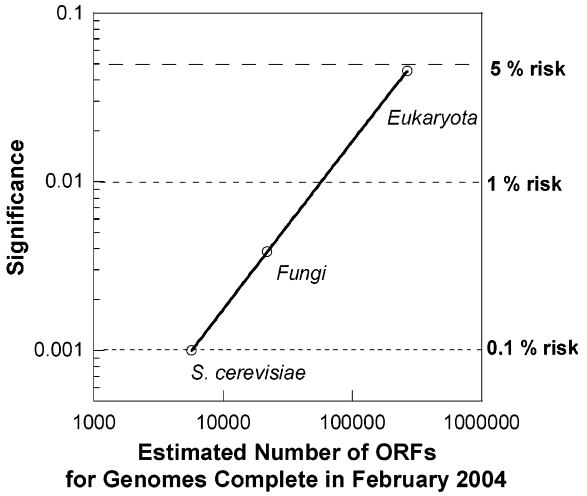

Figure 3. Significance as obtained from eq 6 of a protein identification result of *S. cerevisiae* as a function of searching with different taxonomic precision, i.e., searching a different number of protein sequences. The result with the highest taxonomic precision, i.e., *S. cerevisiae*, is displayed at the 0.1% significance level. The line is a guide for the eye.

Discussion

Similarities of Different Distributions. The function employed for *S in eq 6 is a close relative of an exponential distribution. It can be shown that*

$$S = 1 - (1 - a\beta)^N \approx 1 - e^{-aN\beta} \qquad (7)$$

for small values of β *and large values of* N, and least-squares fits to our data of Figure 1 yield the same values of the parameter a for both functions of eq 7. The right side of eq 7 can be interpreted as a Weibull distribution with the shape parameter equal to 1. The Weibull distribution is an extreme value distribution, but with the observed variable limited by 0 and infinity. The nature of the ranking in Probity and in most other protein identification algorithms is to pick out an extreme value (smallest or largest). The good fitting capability of the functions employed here in the eqs 6 and 7 is a natural consequence of the ranking by an extreme value.

Effects of Sequence Similarities. The scaling of the difficulty of obtaining statistically significant protein identification results as displayed in the Figures 1–3 appears robust over a large range of sizes of the sequence collections. The assumption underlying the computations is that the sequences in a collection are different. The effect on the statistical significance of sequence *similarities* between different sequences in a sequence collection is yet unknown. Using a simple routine that considers sequences that share at least five tryptic peptides as sequence similar revealed that the fraction of the sequences considered as similar ranges broadly from 2% (*H. influenzae*) to 12% (*C. elegans*). We speculate that the effect of sequence similarity is that the effective size, N, of a sequence collection should be somewhat reduced when predicting the significance from the score (eq 6) and therefore significance values as computed here using eq 6 could be somewhat conservative.

Taxonomic Precision. The loss of significance when searching more sequences than necessary (Figure 3) must be interpreted with some caution. We believe that the result indicated by Figure 3 is generally valid for eukaryotes. However, *microbial* genomes can evolve rapidly as a result of environmentally induced selective pressure. Microbial evolution can occur via horizontal gene transfer.[26,27] Therefore, new genes evolved can display high similarity with genes of other microbial organisms. Hence, for microbial proteome analysis it cannot be excluded that searching with less taxonomic precision can indeed improve the quality of identification results simply because a gene expressed might actually not be found in the expected sequence collection but in a sequence collection originating from another organism.

Protein Identification in Higher Organisms. Our simulations reveal explicitly the increased difficulty for statistically significant protein identification when moving from model organisms such as yeast to higher organisms such as human. The practical consequences of the increased difficulty can be understood by the following example: Assume that a laboratory operates with a desired significance level, S', for an organism having N_1 sequences. S' can be viewed as the risk the laboratory is willing to take that a result presented is false. The lower the risk the higher the number of results discarded in a testing procedure. The corresponding significance level, S, if the laboratory keeps the same experimental procedures, but instead studies an organism with N_2 sequences ($N_2 > N_1$) can be computed using eq 6 (Figure 2). If the laboratory still desires the level S' it is expected that the fraction of the results that must be discarded will increase.

An increased number of different proteins in the cells analyzed implies more challenging experimental work, e.g., concerning separation, in addition to the strictly theoretical difficulties elucidated here. High throughput and quality assessment are keys to the successful proteome analysis. A statistically justified approach to discard poor quality results is an important feature of an automated, high throughput system. We believe that the experimental and theoretical aspects of protein identification are intertwined. A more detailed knowledge of the informatics aspect of protein identification should serve as a guide for the optimization of the laboratory work in order to maximize the throughput of significant results for the respective organism studied.

Conclusions

We have established a simple analytical formula that can predict the statistical significance of a protein identification result for any size of the sequence collection searched. We have elucidated the increased difficulty of obtaining significant results in higher organisms by comparing identification in higher organisms with identification in model organisms and also elucidated the loss of significance resulting from searching with suboptimized taxonomic precision.

Acknowledgment. Financial support to J.E. was received from the Swedish Research Council (VR).

References

(1) Fleischmann, R. D.; Adams, M. D.; White, O.; Clayton, R. A.; Kirkness, E. F.; Kerlavage, A. R.; Bult, C. J.; Tomb, J. F.; Dougherty, B. A.; Merrick, J. M.; et al. *Science* **1995**, *269*, 496.
(2) 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.
(3) The C. elegans sequencing consortium. *Science* **1998**, *282*, 2012.
(4) The human genome project. *Nature* **2001**, *409*, 813.
(5) Neubauer, G.; Gottschalk, A.; Fabrizio, P.; Seraphin, B.; Luhrmann, R.; Mann, M. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 385.
(6) Shevchenko, A.; Keller, P.; Scheiffele, P.; Mann, M.; Simons, K. *Electrophoresis* **1997**, *18*, 2591.
(7) Wigge, P. A.; Jensen, O. N.; Holmes, S.; Soues, S.; Mann, M.; Kilmartin, J. V. *J. Cell. Biol.* **1998**, *141*, 967.
(8) Rout, M. P.; Aitchison, J. D.; Suprapto, A.; Hjertaas, K.; Zhao, Y.; Chait, B. T. *J. Cell. Biol.* **2000**, *148*, 635.
(9) Gavin, A. C.; Bosche, M.; Krause, R.; Grandi, P.; Marzioch, M.; Bauer, A.; Schultz, J.; Rick, J. M.; Michon, A. M.; Cruciat, C. M.; Remor, M.; Hofert, C.; Schelder, M.; Brajenovic, M.; Ruffner, H.; Merino, A.; Klein, K.; Hudak, M.; Dickson, D.; Rudi, T.; Gnau, V.; Bauch, A.; Bastuck, S.; Huhse, B.; Leutwein, C.; Heurtier, M. A.; Copley, R. R.; Edelmann, A.; Querfurth, E.; Rybin, V.; Drewes, G.; Raida, M.; Bouwmeester, T.; Bork, P.; Seraphin, B.; Kuster, B.; Neubauer, G.; Superti-Furga, G. *Nature* **2002**, *415*, 141.
(10) Ho, Y.; Gruhler, A.; Heilbut, A.; et al. *Nature* **2002**, *415*, 180.
(11) McLachlin, D. T.; Chait, B. T. *Curr. Opin. Chem. Biol.* **2001**, *5*, 591.
(12) Oda, Y.; Nagasu, T.; Chait, B. T. *Nat. Biotechnol.* **2001**, *19*, 379.
(13) Zhou, H.; Watts, J. D.; Aebersold, R. *Nat. Biotechnol.* **2001**, *19*, 375.
(14) Zhang, H.; Li, X. J.; Martin, D. B.; Aebersold, R. *Nat. Biotechnol.* **2003**, *21*, 660.
(15) Mann, M.; Jensen, O. N. *Nat. Biotechnol.* **2003**, *21*, 255.
(16) 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.
(17) James, P.; Quadroni, M.; Carafoli, E.; Gonnet, G. *Biochem Biophys Res. Commun.* **1993**, *195*, 58.
(18) Mann, M.; Hojrup, P.; Roepstorff, P. *Biol. Mass Spectrom.* **1993**, *22*, 338.
(19) Pappin, D. J. C.; Hojrup, P.; Bleasby, A. *Curr. Biol.* **1993**, *3*, 327.
(20) Yates, J. R. d.; Speicher, S.; Griffin, P. R.; Hunkapiller, T. *Anal. Biochem.* **1993**, *214*, 397.
(21) Zhang, W.; Chait, B. T. *Anal. Chem.* **2000**, *72*, 2482.
(22) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551.
(23) Eriksson, J.; Chait, B. T.; Fenyo, D. *Anal. Chem.* **2000**, *72*, 999.
(24) Eriksson, J.; Fenyo, D. *J. Proteome Res.* **2004**, *3*, 32.
(25) Eriksson, J.; Fenyo, D. *Proteomics* **2002**, *2*, 262.
(26) Jain, R.; Rivera, M. C.; Moore, J. E.; Lake, J. A. *Theor. Popul. Biol.* **2002**, *61*, 489.
(27) Jain, R.; Rivera, M. C.; Moore, J. E.; Lake, J. A. *Mol. Biol. Evol.* **2003**, *20*, 1598.

PR0499343