Anal. Chem. **2000,** *72,* 2482−2489

ProFound: An Expert System for Protein Identification Using Mass Spectrometric Peptide Mapping Information

Wenzhu Zhang* and Brian T. Chait*

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

We describe the protein search engine "ProFound", which employs a Bayesian algorithm to identify proteins from protein databases using mass spectrometric peptide mapping data. The algorithm ranks protein candidates by taking into account individual properties of each protein in the database as well as other information relevant to the peptide mapping experiment. The program consistently identifies the correct protein(s) even when the data quality is relatively low or when the sample consists of a simple mixture of proteins. Illustrative examples of protein identifications are provided.

The rapid expansion of protein and DNA sequence databases together with technological improvements in biological mass spectrometry (MS) has made the combination of mass spectrometric peptide mapping with database searching[1−5] a superb method for rapid protein identification. The method (Figure 1) involves cleavage of proteins with an enzyme having high specificity (usually trypsin), whereupon the resulting proteolytic products are subjected to analysis by either matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS) or electrospray ionization mass spectrometry (ESI-MS). Through the use of an appropriate computer algorithm, the masses determined for the proteolytic peptides are compared with masses calculated for theoretically possible enzymatic cleavage products for every sequence in a protein/DNA sequence database. The protein is identified based on an evaluation of this comparison. This peptide mapping method for protein identification is fast because the mass spectra are rapidly collected (<1 min/spectrum for MALDI-time-of-flight analysis) and because the analysis can be performed on the same time scale. The method is relatively insensitive to unspecified modifications and/or sequence errors in the database because high-confidence identifications can be made even when the mapping experiment yields information on only a small percentage of the sequence.

Identification of proteins by the above-described approach requires a scheme for determining the best match between the experimental data and a sequence in the database. Existing schemes for determining the best match include ranking by number of matches[1−4] and a scoring system based on the observed frequency of peptides from all proteins in a database in a given molecular weight range (the so-called "MOWSE score"[5]). When the mass spectral data are incomplete (i.e., only a few peaks in the spectrum) and/or of low mass accuracy, the "number-of-matches" approach may be inadequate to make a useful identification. While the MOWSE scoring scheme is much superior to the number-of-matches approach, it does not take into account the individual properties of any given protein. An optimal scoring system requires that individual properties of each protein in the database be considered.

In the present paper, we describe an expert system for identifying proteins using MS peptide mapping data. The system ranks protein candidates using a Bayesian algorithm that takes into account individual properties of each protein in the database as well as other information relevant to the experiment. Bayesian probability theory has been widely used to make scientific inference from incomplete information in various disciplines, including biopolymer sequence alignment,[6] NMR spectral analysis,[7] and radar target identification.[8] When the system under study is modeled properly, the Bayesian approach is believed to be always among the most coherent, consistent, and efficient statistical methods.[6−10] Here, we apply Bayesian probability theory to make logical inference about the identity of an unknown protein sample against a protein sequence database. The probability for the sample protein to be a specific protein in the database is calculated using the MS data as well as other background information such as the mass range in which the protein is expected to lie, identity of species from which the protein originated, mass accuracy, enzyme cleavage chemistry, protein sequence, previous experiments on the sample protein, etc. A first

* Corresponding authors: (e-mail) chait@rockvax.rockefeller.edu.; zhangw@rockvax.rockefeller.edu.

(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) Yates, J. R., III; Speicher, S.; Griffin, P. R.; Hunkapiller, T. *Anal. Biochem.* **1993,** *214,* 397−408.
(3) Mann, M.; Hojrup, P.; Roepstorff, P. *Anal. Chem.* **1993,** *22,* 338−45.
(4) James, P.; Quadroni, M.; Carafoli, E.; Gonnet, G. *Biochem. Biophys. Res. Commun.* **1993,** *195,* 58−64.
(5) Pappin, D. J. C.; Hojrup, P.; Bleasby, A. J. *Curr. Biol.* **1993,** *3,* 327−32.
(6) Liu, J. S.; Lawrence, C. E. *Bioinformatics* **1999,** *15,* 38−52.
(7) Bretthorst, G. L. *Bayesian spectrum analysis and parameter estimation*; Lecture Notes in Statistics 48; Springer-Verlag: New York, 1988.
(8) Bretthorst, G. L. In *Maximum Entropy and Bayesian Methods*; Heidbreder, G. R., Ed.; Kluwer Academic: Dordrecht, 1996; pp 1−42.
(9) Jaynes, E. T. *Probability theory: The logic of science*; Cambridge University Press: to be published.
(10) Gelman, A.; Carlin, J. B.; Stern, H. S.; Rubin, D. B. *Bayesian data analysis*; Chapman & Hall: New York, 1995.

10.1021/ac991363o CCC: $19.00
Published on Web 05/04/2000

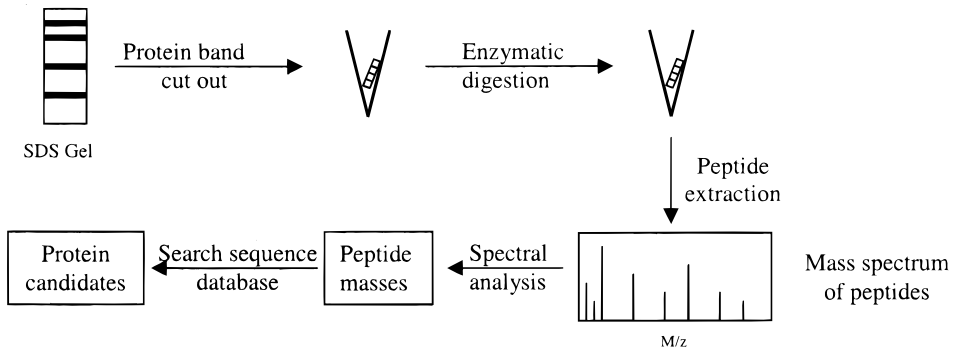

Figure 1. Flowchart showing the procedure for mass spectrometric protein identification.

version of the algorithm was implemented in 1995[11] and made publicly available over the World Wide Web as a protein identification tool called ProFound (URL http://prowl.rockefeller.edu/cgi-bin/ProFound). Up until the present, the program has been accessed >150 000 times, during which time we have had extensive feedback from users and numerous requests for details concerning the algorithm. The algorithm and the program have been continually improved and extended to incorporate data from multiple digestions, to utilize additional experimental information concerning the amino acid content of individual peptides, and to identify protein components in mixtures.[12] In this paper, we provide a detailed description of the current Bayesian algorithm and the program together with examples illustrating its use for protein identification.

METHODS

Algorithm. Every protein is specified by its particular linear sequence of amino acids. One defining signature of a protein is the set of masses of peptide fragments produced by cleavage of the protein by an enzyme of high cleavage specificity. The problem we seek to solve is to use the peptide masses obtained in such a mass spectrometric peptide mapping experiment to identify a protein from a protein sequence database.

Let k designate the hypothesis that "protein k is the protein being analyzed", where protein k is an entry in the protein sequence database, D is the experimental data, and I is the available background information (e.g., species from which the protein originated, approximate molecular mass of the protein, mass accuracy of the peptide mass measurement, enzyme cleavage chemistry, previous experiments on the sample protein, etc.). Bayes' probability theory and the maximum entropy principle[8] are applied to derive the probability for the hypothesis k given data D and background information I (see Supporting Information). In the derivation, the following assumptions are made: (1) the protein being analyzed exists in the database; (2) all the detected ion species are digestion products of the protein; (3) when a hit (a match between a measured mass and a calculated theoretical peptide mass within given mass accuracy) occurs, the measured peptide is regarded as the theoretical peptide (i.e., the random hit situation is not considered). The probability for each hypothesis k given data D and background information I is given by (see Supporting Information for the derivation)

$$P(k|DI) \propto P(k|I)\,\frac{(N-r)!}{N!}\prod_{i=1}^{r}\left\{\sqrt{\frac{2}{\pi}}\,\frac{m_{\max}-m_{\min}}{\sigma_i}\times \sum_{j=1}^{g_i}\exp\left[-\frac{(m_i-m_{ij0})^2}{2\sigma_i^{\,2}}\right]\right\}F_{\text{pattern}} \quad (1)$$

with the normalization condition

$$\sum_{k\,\in\text{database}} P(k|DI) = 1 \quad (2)$$

The ranking of the candidate proteins is based on the values determined for their probability $P(k|DI)$. $P(k|I)$ in eq 1 is the probability for hypothesis k given only the background information, I; N is the theoretical number of peptides generated by fragmentation of protein k by the protease used in the study; r is the number of hits (i.e., the number of matches between the measured and calculated peptide masses); $(m_{\max} - m_{\min})$ is the range of measured peptide masses; m_i is the measured mass of the ith hit, which has multiplicity of g_i (i.e., the number of theoretical peptides that match m_i); m_{ij0} is the calculated mass of the jth peptide in the ith hit; σ_i is the standard deviation of the mass measurement at mass m_i; and F_{pattern} is an empirical term, which increases the probability when overlapping and/or adjacent peptides are observed (see program description). For large N, eq 1 can be approximated as,

$$P(k|DI) \sim P(k|I)\left(\sqrt{\frac{2}{\pi}}\,\frac{m_{\max}-m_{\min}}{N}\right)^{r}\times \prod_{i=1}^{r}\frac{1}{\sigma_i}\left\{\sum_{j=1}^{g_i}\exp\left[-\frac{(m_i-m_{ij0})^2}{2\sigma_i^{\,2}}\right]\right\}F_{\text{pattern}} \quad (3)$$

Equation 3 is useful for interpreting the effect of the various terms on $P(k|DI)$.

Interpretation of the Probability $P(k|DI)$. The Bayesian probability is consistent with common sense. For any given protein k in the database, the probability that protein k is the sample protein increases with increasing number of hits r, increasing mass

(11) Zhang, W.; Chait, B. T. *Proceedings of the 43rd ASMS Conference on Mass Spectrometry and Allied Topics*, Atlanta, GA, 1995; p 643.
(12) Zhang, W.; Chait, B. T. *Proceedings of the 46th ASMS Conference on Mass Spectrometry and Allied Topics*, Orlando, FL, 1998; p 969.

accuracy (i.e., smaller σ_i and $(m_i - m_{ij0})$), and decreasing number of theoretically digested fragments N (see eq 3).

The Bayesian probability should be viewed as a measure of the confidence level of the hypothesis that protein k is the sample protein based on the available information. There is no absolute certainty for any given identification, only the probability—i.e., the higher the probability, the higher is the confidence level. In a situation where a false positive result cannot be tolerated or the data are of insufficiently quality to yield a probability with a high level of discrimination, it is desirable to check the identification with an independent method such as tandem mass spectrometry (MS/MS).[13−15] Ultimately, the value of any given identification is provided by the outcome of the biological experiment that results from the information.

Identification of Components in Protein Mixtures. Frequently, it proves difficult to separate proteins completely from one another, and a protein sample may contain a mixture of proteins. The Bayesian algorithm can be readily extended to identify the components of such mixtures. The protein sequence database is expanded to include entries that are "fused" combinations of single protein sequences. An early version of the program used this approach to identify the components of binary mixtures.[12] Thus, the entries representing binary mixtures are binary combinations of single proteins (usually, the top 50 hits obtained in a prior search for single proteins). The Bayesian probabilities for these "fused" proteins are calculated in the same way as for single proteins. Using this fused-protein approach, a current version of the search engine (also available over the World Wide Web) allows identification of up to four protein components in a mixture.

Improvement of the Confidence Level of Protein Identification Using Additional Information Obtained for the Measured Peptides. The Bayesian algorithm can incorporate any additional information obtained for measured peptides (see Supporting Information). The additional information provides constraints in database searching to reduce the occurrence of database peptides that randomly match the experimental mass spectral data, thereby improving the confidence level for identifications. Fenyö et al. have investigated the value provided by the knowledge of the presence (or absence) and the number of particular amino acids contained within a given peptide (so-called "tag information").[16] Experimentally, tag information can be obtained in a number of ways. Thus, for example, cysteine residues can be identified through chemical alkylation of free thiol moieties[17] and methionine residues can be inferred by observation of pairs of peaks separated by 16 Da (because methionine residues contained in proteolytic peptides are frequently found to be partially oxidized).

Program. ProFound is publicly available over the Internet at URL http://prowl.rockefeller.edu as part of PROWL—An interactive environment on the World Wide Web for protein MS.[18] The database searches were performed on an Origin 200 (2× RA 10000 processors) computer (Silicon Graphics Inc.) and a PC (2× 550 MHz Pentium III Xeon processors) (Dell Computer Corp.). Search times are dependent on the number of masses in the peptide map and the constraint placed on the database. For example, the typical search time for the identification of a yeast protein is ∼1 s on the Pentium III computer. The database searched is the NCBI NR nonredundant database (URL http://www.ncbi.nlm.nih.gov/BLAST/blast_databases.html). The presence or absence of signal peptides is considered when such information is available in the corresponding NCBI GenPept format flatfile (URL http://www.ncbi.nlm.nih.gov/Entrez/batch.html). Taxonomy data are derived with high accuracy from the NCBI GenPept flatfiles and Taxonomy databases (URL http://www.ncbi.nlm.nih.gov/Taxonomy/).

Because there is a ∼95% probability for Gaussian-distributed measurement errors to be within $\pm 2\sigma$ (where σ is the standard deviation), the *mass tolerance* is taken as 2σ in the probability calculation. Mass-independent systematic errors in the mass measurements have been removed during the probability calculations (see text for eq A10 in Supporting Information). An empirical factor has also been introduced in the probability calculation to take into account two kinds of commonly observed digestion patterns. The first pattern, which we term adjacency, occurs when proteolytic peptides are observed to be adjacent to one another in the protein sequence (see Figure 4A). The second pattern, which we term common-end overlapping, occurs when the observed peptides have one common terminus but differ at the other terminus by a peptide segment (see Figure 4A). The probability is increased on each occurrence of adjacency or common-end overlapping.

The following are the current ProFound input parameters: *taxonomy category*, specifying the origin of the sample protein, if known, from a representation of a phylogenic tree; *mass range*, specifying the approximate protein mass range of the sample protein, if known; *digestion chemistry*, specifying the proteolytic enzyme or chemical reagent used to cleave the sample protein(s); *maximum number of missed cleavage sites*, specifying the maximum number of missed cleavage sites within the peptide (yielding incompletely cleaved peptides); *modifications*, modifications of amino acid residues can be specified..

To obtain the best results from ProFound, it is necessary to choose the optimum search parameters for a particular set of experimental data. Thus, for example, if the digest is carried out to completion, it may be prudent to set the maximum number of missed cleavages at 0 or 1. On the other hand, if the digest is rather incomplete it may prove advantageous to set the maximum number of missed cleavages at >1.

Biochemical Procedures and Mass Spectrometry. The method used for in-gel tryptic digestion of proteins was as described[19] except that the gel-band washing time was extended from 4 to 24 h. Endoproteinase LysC digestion of membrane-bound proteins was as described previously.[20] MALDI-time-of-flight (TOF) MS was carried out using a commercial instrument (Perseptive Biosystems STR, Framington, MA) operated in the delayed-extraction reflector mode (fwhm resolution ∼5000) or an

(13) Yates, J. R., III *Electrophoresis* **1998,** *19*, 893−900.
(14) Kuster, B.; Mann, M. *Curr. Opin. Struct. Biol.* **1998,** *8*, 398−400.
(15) Patterson, S. D.; Aebersold, R. *Electrophoresis* **1995,** *16*, 1791−814.
(16) Fenyö, D.; Qin, J.; Chait, B. T. *Electroporesis* **1998,** *19*, 998−1005.
(17) Sechi, S.; Chait, B. T. *Anal. Chem.* **1998,** *70*, 5150−8.
(18) Fenyö, D.; Zhang, W.; Beavis, R.; Chait, B. T. *Anal. Chem.* **1996,** *68*, A721.
(19) Zhang, X.; Herring, C. J.; Romano, P. R.; Szczepanowska, J.; Brzeska, H.; Hinnebusch, A. G.; Qin, J. *Anal. Chem.* **1998,** *70*, 2050−9.
(20) Zhang, W.; Czernik, A. J.; Yungwirth, T.; Aebersold, R.; Chait, B. T. *Protein Sci.* **1994,** *3*, 677−86.

Table 1. ProFound Search Results Obtained with the Data Shown in Figure 2

rank	probability	gene name[a]	description	mass (kDa)
1	1	*rps4b*	ribosomal protein S4	29
2	2×10^{-51}	*myo1*	myosin heavy chain type II	214
3	8×10^{-53}	*nup82*	nuclear pore protein nup82	82
4	5×10^{-53}	*dyn1*	cytoplasmic dynein heavy chain	471

[a]Yeast Protein Database gene name.

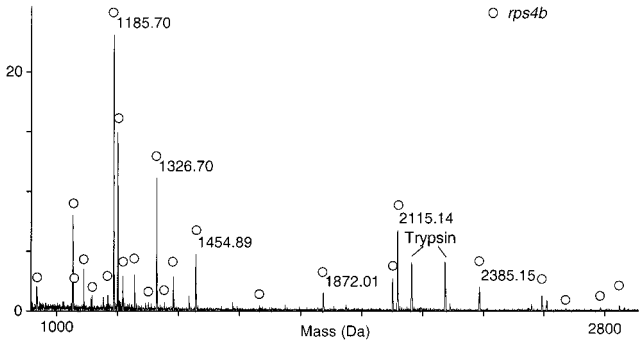

Figure 2. Delayed-extraction reflectron MALDI-TOF spectrum of the proteolytic products from an in-gel tryptic digest of a 30 kDa SDS−PAGE protein band. The search engine, ProFound, identified a single protein: *rps4b* (40s ribosomal protein S4). Open circles indicate peaks that match with masses of the theoretical tryptic fragments of *rps4b*. Trypsin self-digestion products are labeled "Trypsin".

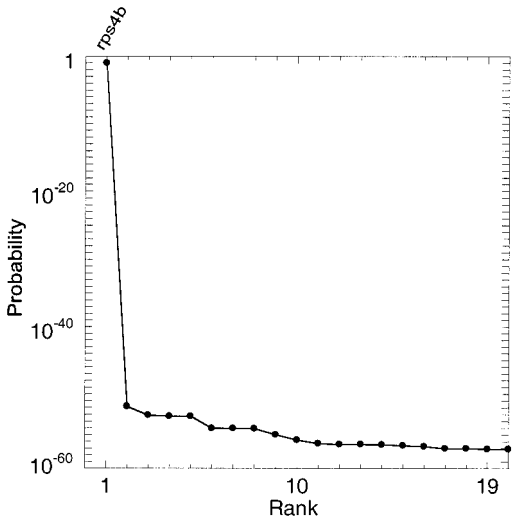

Figure 3. Normalized probability distribution for the top 20 protein candidates identified by ProFound using the data shown in Figure 2.

instrument constructed in-house[21] operated in the continuous-extraction linear mode (fwhm resolution ~500). The MALDI-ion trap data were obtained using an instrument constructed in-house and described previously.[22]

RESULTS AND DISCUSSION

Identification of Single Isolated Proteins. Figure 2 shows a delayed-extraction reflectron MALDI-TOF spectrum of the mixture of peptides produced by in-gel trypsin digestion of a 30 kDa SDS−PAGE protein band from a *Saccharomyces cerevisiae* nuclear extract. Thirty-five monoisotopic masses derived from Figure 2 were submitted to ProFound in order to identify the protein. Other search parameters were as follows: *S. cerevisiae* for the taxonomic category; a protein mass range of 0−3000 kDa; unmodified cysteines; a maximum of two missed cleavage sites; a mass tolerance of 0.1 Da. The specified taxonomic category and protein mass range includes the complete set of proteins (or open reading frames (ORFs)) in the *S. cerevisiae* genome. Table 1 lists the top four protein candidates (ranked by normalized probability) found by the search. The top-ranked protein, *rps4b*, has a probability very close to 1 and is readily distinguished from the next ranked candidates, which have probabilities of respectively 2×10^{-51}, 8×10^{-53}, and 5×10^{-53}. We plot the probabilities of the top 20 candidates in Figure 3. The probability is observed to make a large transition from the first to second candidate and varies much more slowly for the remaining candidates. This type of probability distribution pattern provides an unambiguous high confidence identification signature for the top ranked protein.

Parts A−C of Figure 4 shows sequence coverage maps and an error map for the top ranked candidate. The segment coverage map (Figure 4A) (in which a segment represents a peptide resulting from complete digestion of the protein by trypsin) is useful for visualizing digestion patterns indicative of an authentic protein identification. Bona fide identifications are often characterized by the observation of peptides that are adjacent to one another in the sequence and/or that overlap and have a common terminal (while differing by one segment at the other terminal.) Examples of these two commonly observed patterns are shown in Figure 4A. Because the observation of such patterns raises our confidence level that a candidate protein is present in the sample, we have empirically included a term in the ProFound probability calculation to incorporate this information. The sequence residue coverage map (Figure 4B) shows the portion of *rps4b* sequence that was observed in the MS peptide mapping experiment. Twenty-three measured masses match 24 theoretical tryptic peptide masses from *rps4b*, covering 70% of the sequence. The error map (Figure 4C) provides a scatter plot showing error (i.e., measured mass − calculated mass) versus mass for each match. The scatter plot is useful for visualizing systematic errors in the mass measurement. When the spectral calibration is free of systematic error, the errors for an authentic hit are normally distributed about zero and are independent of mass value, as in Figure 4C. The bottom portion of Figure 4C is a histogram projection of the scatter plot. In cases where there are a sufficient number of matched peaks, the histogram of errors for an authentic hit shows a peaked distribution (Figure 4C). By contrast, the error plot for a randomly hit protein (e.g., the second candidate, *myo1*, Figure 4D) is a nearly uniform distribution of mass errors within the plotted range. We

(21) Beavis, R. C.; Chait, B. T. *Rapid Commun. Mass Spectrom.* **1989,** *2*, 233−7.
(22) Qin, J.; Steenvoorden, R. J. J. M.; Chait, B. T. *Anal. Chem.* **1996,** *68*, 1784−91.

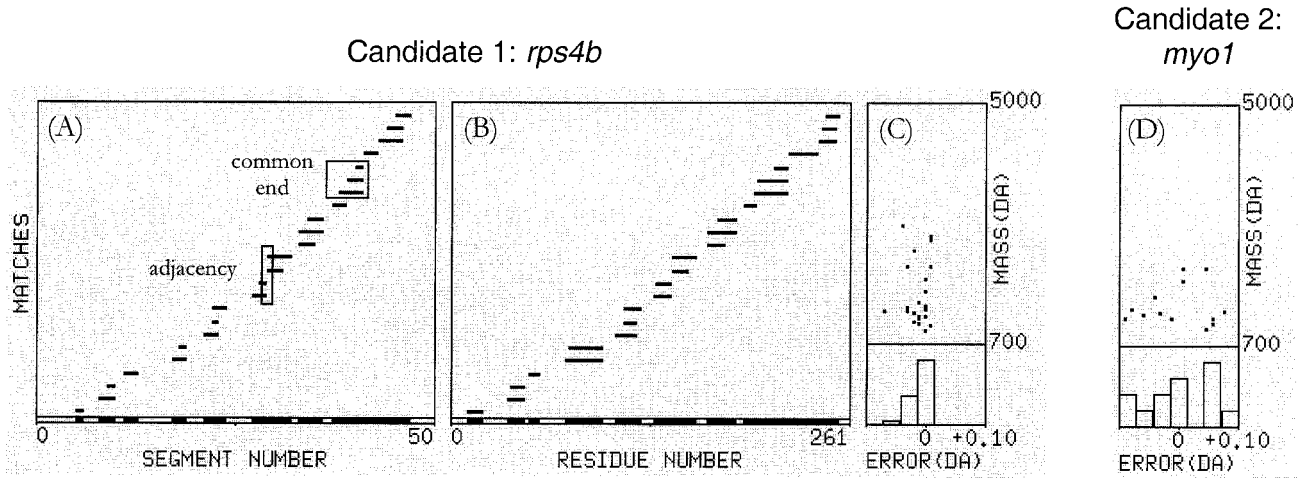

Figure 4. Graphical representation of the sequence coverage of the top-matched protein by measured peptides. (A) and (B) are respectively the segment and residue coverage maps (see text) for the top candidate, *rps4b,* identified by ProFound based on the data shown in Figure 2. (C) and (D) are error maps for, respectively, *rps4b* and the second most probable protein candidate, *myo1*.

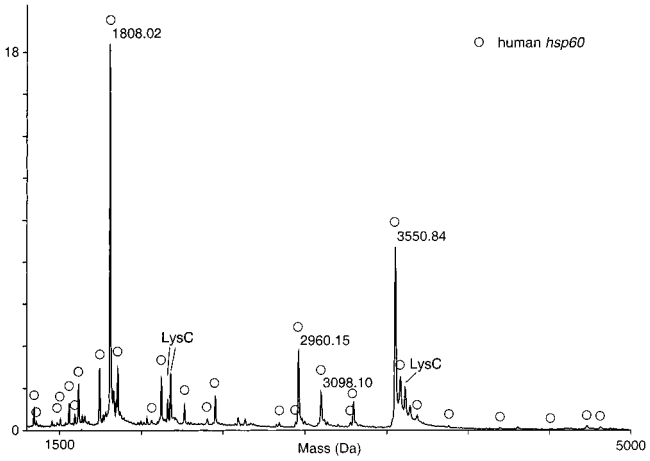

Figure 5. Direct-extraction linear MALDI-TOF spectrum of endoproteinase LysC digest of a protein on a modified PVDF membrane blotted from a 2Dgel. ProFound determined that the band was a single protein: human *hsp60* (human heat shock protein 60). Open circles indicate peaks that match with masses of theoretical LysC proteolytic fragments of human *hsp60*. Endoproteinase LysC self-digestion products are labeled "LysC".

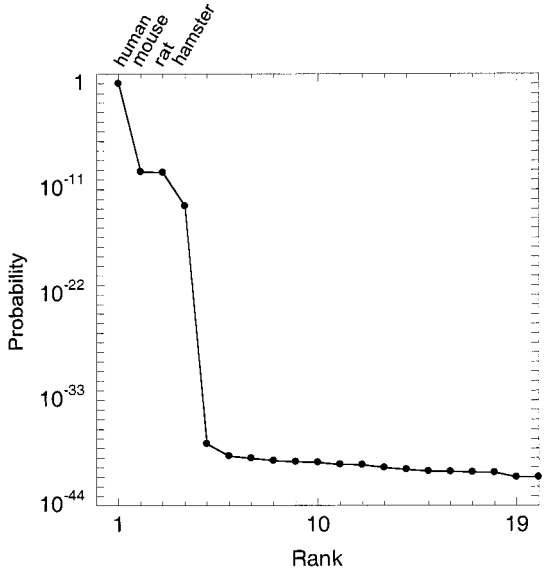

Figure 6. Normalized probability distribution for the top 20 protein candidates identified by ProFound using the data shown in Figure 5.

note, however, that this observation only holds true when the error of the mass measurement is small (see Supporting Information) and the mass tolerance is properly chosen.

Figure 5 provides an example of a relatively low resolution MALDI-TOF spectrum taken with a linear TOF analyzer operated in the continuous-extraction mode. The sample protein was obtained from a human mitochondrial preparation and was blotted to a membrane after two-dimensional electrophoretic separation. The protein spot was digested on the membrane with endoproteinase LysC, and the mixture of proteolytic peptides was extracted and subjected to MS. The resulting mass spectrum yielded average masses of 36 peptides, which were submitted to ProFound for protein identification. Other search parameters were as follows: *all taxa* for taxonomic category; a protein mass range of 0−3000 kDa; cysteines modified by acrylamide; a maximum of four missed cleavage sites; a relative mass accuracy tolerance of 0.04%. Table 2 shows results of the search and Figure 6 the probability plot as a function of the protein candidates. The resulting pattern of probabilities (Figure 6) is different from that seen in the previous example (Figure 3). Although there is a clear preference for human heat shock protein 60 ($P \approx 1$), several highly homologous heat shock proteins from other species yield considerably higher probabilities ($P = 7 \times 10^{-10}$, 6×10^{-10}, and 2×10^{-13}) than those observed for the remaining randomly hit proteins ($P < 10^{-37}$). The sequence coverage maps for the first- and second-ranked candidates (human *hsp60* and its mouse homologue) exhibit similar patterns (maps not shown), with, respectively, 29 and 25 peak matches with peptide masses from the human and mouse proteins.

Identification of Protein Components In Mixtures. The MALDI-TOF mass spectrum shown in Figure 7 was obtained from the products of in-gel trypsin digestion of a 105 kDa SDS−PAGE protein band. Peptide masses consisting of 47 monoisotopic masses were submitted to ProFound. Other search parameters were as follows: *S. cerevisiae* as taxonomic category; protein mass

Table 2. ProFound Search Results Obtained with the Data Shown in Figure 5

rank	probability	gene name	description	mass (kDa)
1	1	human *hsp60*	human heat shock protein 60	58
2	7×10^{-10}	mouse *hsp60*	mouse heat shock protein 60	61
3	6×10^{-10}	rat *hsp60*	rat heat shock protein 60	58
4	2×10^{-13}	chinese hamster *hsp60*	chinese hamster heat shock protein 60	58
5	3×10^{-38}	*Dictyostelium discoideum acaa*	*D. discoideum* adenylate cyclase	160

Table 3. ProFound Search Results with Data Obtained from In-Gel Tryptic Digest Shown in Figure 7

rank	probability	gene name	description	mass (kDa)
1	1	*YLR409c*	protein of unknown function	105
		YDL060w	protein of unknown function	91
2	2×10^{-23}	*YLR409c*	protein of unknown function	105
		YNR051c	protein with weak similarity to chicken nucleolin, has an RNA recognition domain	58
3	5×10^{-26}	*YLR409c*	protein of unknown function	105
		cdc39	nuclear protein that negatively affects basal transcription from many promoters	240

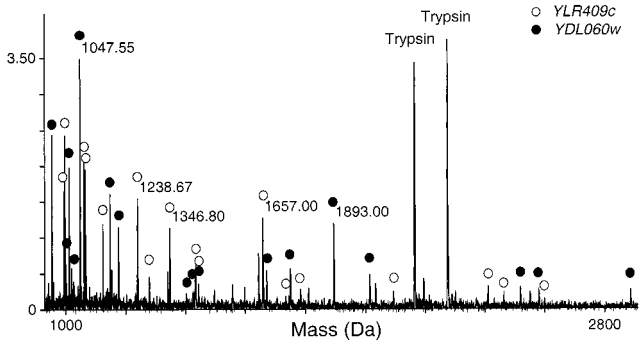

Figure 7. Delayed-extraction reflectron MALDI-TOF spectrum of an in-gel tryptic digest of a 105 kDa SDS−PAGE protein band. ProFound determined that the band was a mixture, identifying two protein components: *YLR409c* and *YDL060w*. Open and solid circles indicate peaks that match with masses of theoretical tryptic fragments of *YLR409c* and *YDL060w*, respectively. Trypsin self-digestion products are labeled "Trypsin".

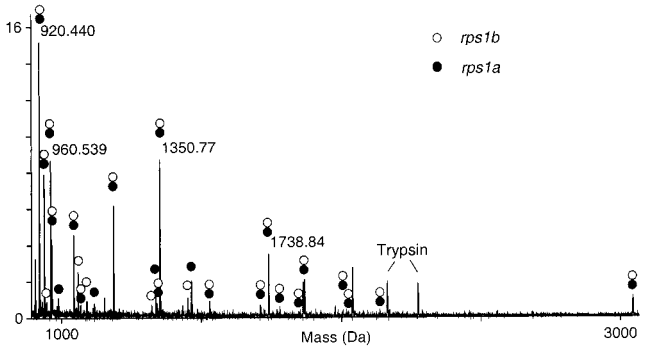

Figure 8. Delayed-extraction reflectron MALDI-TOF spectrum of an in-gel tryptic digest of a 30 kDa SDS−PAGE protein band. ProFound determined that the band was a mixture, identifying two protein components: *rps1B* and *rps1A*. Open and solid circles indicate peaks that match with masses of theoretical tryptic fragments of *rps1B* and *rps1A*, respectively. Trypsin self-digestion products are labeled "Trypsin".

range of 0−3000 kDa; unmodified cysteines; two maximum missed cleavage sites; a mass tolerances of 0.2 Da. When the search was performed in the "single protein" mode, the top two candidates (*YLR409c* and *YDL060w*) had probabilities of 0.99 and 0.01, respectively, which was considerably higher than the probabilities for all the rest of the candidate proteins ($\leq 10^{-22}$ with slowly decreasing values). The number of peaks matching with theoretical tryptic peptide masses from *YLR409c* and *YDL060w* are 18 each (respective sequence coverage of 24% and 30%), and the two proteins have no sequence homology. Two such proteins with dominating probabilities provide an indication that the sample may be a binary mixture. To test this hypothesis, ProFound was set up to search for a possible binary mixture using the same data set and search parameters that were used for the "single protein only" search. The result of this search identifies, with high confidence, the simultaneous presence of *YLR409c* and *YDL060w* (Table 3). This binary protein mixture has a probability very close to unity, while the probabilities for all the other protein candidates are $\leq 10^{-23}$, with slowly decreased values. The validity of this mixture identification can be tested by the use of the subtraction method.[23] For this purpose, the 18 peaks corresponding to the tryptic peptides from the highest probability component (*YLR409c*) were removed from the peptide mass list. The remaining masses, together with the same search parameters, were submitted to ProFound to search for the hypothetical second protein. ProFound identifies *YDL060w* as the leading candidate with a probability very close to 1 and $\sim 10^{18}$ higher than the next protein candidate.

Figure 8 shows a MALDI-TOF mass spectrum obtained from in-gel tryptic digestion of a 30 kDa SDS−PAGE protein band. Thirty-six monoisotopic peptide masses and one average peptide mass were submitted to ProFound. Other search parameters were as follows: *S. cerevisiae* as taxonomic category; protein mass range of 0−3000 kDa; unmodified cysteines; a maximum of two missed cleavage sites; mass tolerances 0.5 Da for average and 0.1 Da for monoisotopic masses. When the search mode was for "single protein only", the top two candidates are *rps1b* (40S ribosomal protein, MM = 28 681 Da) and *rps1a* (40S ribosomal protein, MM = 28 612 Da, highly homologous to *rps1b*) with probabilities of, respectively, 1 and 0.001, while the probability for the remaining

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

Table 4. ProFound Search Results with Data Obtained from In-Gel Tryptic Digest Shown in Figure 8

rank	probability	gene name	description	mass (kDa)
1	1	*rps1b*	40S ribosomal protein	29
		rps1a	40S ribosomal protein.	29
2	1×10^{-14}	*rps1b*	40S ribosomal protein	29
		YER053c	putative mitochondrial carrier.	34
3	1×10^{-15}	*rps1b*	40S ribosomal protein.	29
4	5×10^{-17}	*rps1a*	40S ribosomal protein	29
		YER053c	putative mitochondrial carrier	12

Table 5. Summary of Identifications Made by Ion Trap MS/MS and Peptide Mapping

		ProFound identification			
gene name[a]	mass (kDa)	rank	$\log(p_1) - \log(p_2)$[b]	match no./total peak no.	sequence coverage (%)
sup35	77	1	14	18/35	43
sup35	77	1	11	16/37	41
sup35	77	1	5	14/30	31
tef1	50	1	4	9/18	36
tef1	50	1	5	11/24	42
ded1	66	1	8	13/24	40
rps14a	15	1	7	8/15	49
ssd1	140	1	16	24/41	44
dbp1	68	1	11	18/36	45
hrr25	57	1	2	8/16	29
ssd1	140	1	4	22/30	29
dbp1	68	1	15	25/46	63

[a] Gene names of the proteins identified by ion trap MS/MS. [b] $\log(p_1) - \log(p_2)$ is the logarithm difference of probabilities between first and second candidate proteins.

candidates are $\leq 10^{-53}$. The number of peaks that match theoretical tryptic peptide masses from *rps1b* and *rps1a* are 24 and 23, respectively (with sequence coverages of 65% and 62%). There are two possibilities that can yield two such dominating candidates. The first is that there is only one protein present in the sample (and the second ranked candidate represents a closely similar protein that is not present). The second possibility is that the sample is a binary protein mixture of two highly similar proteins. To test this second hypothesis, ProFound was set up to search for possible binary mixtures with the same data set and search parameters that were used for the single protein only search. The result of this binary search (Table 4) provides strong evidence that the band is a mixture of *rps1b* and *rps1a* ($P \approx 1$). The probabilities for all the other single and binary protein candidates are $\leq 10^{-14}$, with slowly decreasing probability values. The two identified proteins are highly homologous, differing by only 7 amino acids in their 254-amino acid sequences. Of the 37 peptide peaks in the mass spectrum, 19 are common to both *rps1b* and *rps1a*, while 5 correspond to *rps1b* only, and 4 correspond to *rps1a* only. Unlike the previous example, it would be difficult to identify the second protein component with high confidence using the subtraction method[23] because only 4 peptides (together with 9 peaks that did not match *rps1a*) would remain after subtraction of the 24 peaks from the first protein *rps1b*.

Independent Verification of the MS Peptide Mapping Method for Protein Identification. Two independent strategies for mass spectrometric protein identification are peptide mapping and fragmentation of individual peptide ions (MS/MS).[1−5,13−15] Here, we use MS/MS as a method for independently checking the accuracy of the peptide mapping strategy for identifying proteins.

In a previously described experiment,[24] we used matrix-assisted laser desorption/ionization-ion trap-mass spectrometry to obtain both tryptic map MS data and MS/MS fragment ion data from the same samples. The MS/MS fragmentation information was used to identify proteins with the program PepFrag,[16] which requires an exact match of the peptide mass and the peptide fragment masses with theoretical masses generated from a database-derived peptide. Here, we compare the result obtained with ProFound using the peptide mapping data with the independent identification using MS/MS data from the same sample. Search parameters were as follows: *S. cerevisiae* as taxonomic category; protein mass range of 0−3000 kDa; one missed cleavage site; mass tolerance of 2 Da. Table 5 is a summary of 12 searches using the two independent methods. All the proteins identified with the MS/MS data were confirmed by ProFound using the peptide mapping data, even though the mapping data were of relatively low quality (i.e., resolution 500 fwhm, accuracy ±2 Da). These findings provide independent assurance of the reliability of ProFound for identifying proteins.

Improvement of the Confidence Level of Protein Identification Using Tag Information. Incorporation of amino acid "tag information" in the ProFound search (see Methods) can reduce the occurrence of database peptides that randomly match the experimental MS data, thereby improving the confidence level of an identification. For example, we have shown previously that inclusion of information regarding the absence or presence of cysteine residues in tryptic peptides from proteins can significantly improve the confidence level of a protein identification.[17]

(24) Qin, J.; Fenyö, D.; Zhao, Y.; Hall, W. W.; Chao, D. M.; Wilson, C. J.; Young, R. A.; Chait, B. T. *Anal. Chem.* **1997,** *69*, 3995−4001.

CONCLUSIONS

We have described an expert system, ProFound, which makes optimal use of information derived from MS peptide mapping experiments to identify proteins present in gel bands. The algorithm uses Bayesian theory to rank the candidate proteins by their probability of occurrence and can naturally incorporate additional information about the sample (e.g., additional proteolytic digest information (details discussed only in Supporting Information), amino acid tag information, and sequence information). The algorithm also allows the identification of protein components of mixtures. The present strategy makes explicit use of detailed information concerning the candidate proteins in the database (including the number of theoretically cleaved peptides) and the experimental data (e.g., mass deviations from the masses of theoretical peptide fragments). In addition, ProFound makes empirical use of information concerning fragmentation pattern commonly observed in proteolytic digests. We have observed ProFound to be robust in that it consistently identifies the correct protein even when the MS data quality is relatively low or the protein is present as a component of a simple protein mixture.[25] Although the probability scores calculated by ProFound provide an objective means for identifying proteins, it is highly desirable to also provide a confidence level for assessing the correctness of the identification. We are currently investigating various methods for scoring such confidence levels.[26,27]

(25) Krutchinsky, A. N.; Zhang, W.; Chait, B. T. *J. Am. Soc. Mass Spectrom*. In press.

ACKNOWLEDGMENT

This work was supported by the National Institute of Health (Grant RR00862 from the National Center for Research Resources). W.Z. thanks Professor Phil C. Gregory (University of British Columbia) for his enthusiastic introduction of the Bayesian theory in his lectures. We thank Chao Tang for his important contributions during the whole course of the development of ProFound. We thank Ronald Beavis, Steven Cohen, Jan Eriksson, David Fenyö, Yoshiya Oda, Julio Cesar Padovan, Jun Qin, Salvatore Sechi, Rong Wang, and Yingming Zhao for valuable discussions and contributions to the development of ProFound. Among the people who gave us feedback, we especially thank Farzin Gharahdaghi, Ole Jensen, Scott Patterson, Kathy Stone, and Kenneth Williams for their many suggestions. We also thank Ruedi Aebersold, Tom Yungwirth, and Michael Rout for providing gel-separated proteins.

SUPPORTING INFORMATION AVAILABLE

Appendix 1, derivation of the Bayesian probability that protein *k* is the protein under analysis, and Appendix 2, error distributions of randomly and authentically matched peptides. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review November 29, 1999. Accepted March 24, 2000.

AC991363O

(26) Eriksson, J.; Chait, B. T.; Fenyö, D. *Anal. Chem*. **2000,** *72*, 999−1005.
(27) Tang, C.; Zhang, W.; Fenyö, D.; Chait, B. T. In preparation.