Probability Model for Assessing Proteins Assembled from Peptide Sequences Inferred from Tandem Mass Spectrometry Data

Jian Feng,[†] Daniel Q. Naiman,[†] and Bret Cooper*[,‡]

Department of Applied Mathematics and Statistics, The Johns Hopkins University, Baltimore, Maryland 21218, and Soybean Genomics and Improvement Laboratory, USDA-ARS, 10300 Baltimore Avenue, Building 006, Room 213, Beltsville, Maryland 20705

In shotgun proteomics, tandem mass spectrometry is used to identify peptides derived from proteins. After the peptides are detected, proteins are reassembled via a reference database of protein or gene information. Redundancy and homology between protein records in databases make it challenging to assign peptides to proteins that may or may not be in an experimental sample. Here, a probability model is introduced for determining the likelihood that peptides are correctly assigned to proteins. This model derives consistent probability estimates for assembled proteins. The probability scores make it easier to confidently identify proteins in complex samples and to accurately estimate false-positive rates. The algorithm based on this model is robust in creating protein complements from peptides from bovine protein standards, yeast, *Ustilago maydis* cell lysates, and *Arabidopsis thaliana* leaves. It also eliminates the side effects of redundancy and homology from the reference databases by employing a new concept of peptide grouping and by coherently distinguishing distinct peptides from unique records and shared peptides from homologous proteins. The software that runs the algorithm, called PANORAMICS, provides a tool to help analyze the data based on a researcher's knowledge about the sample. The software operates efficiently and quickly compared to other software platforms.

Tandem mass spectrometry (MS/MS) coupled with multidimensional separations (MudPIT, multidimensional protein identification technology) provides a high-throughput and sensitive way to identify proteins from complex mixtures acquired from cells or tissues. In this method, proteins are digested by a protease, and then the peptides are separated by high-performance liquid chromatography (HPLC), ionized, and sprayed into a mass spectrometer.[1−5] The first mass scan works as a selector which only permits peptides with selected m/z values to pass through. These precursor ions are then fragmented by collision-induced dissociation using an inert gas, resulting in predictable fragmentation along the peptide backbone. A tandem mass spectrum can be analyzed to determine the amino acid sequence of a peptide, which is deduced from the mass losses.[1,6]

Since the tandem mass spectral data sets can be quite large and complex, most scientists rely on computer algorithms to interpret the data. These algorithms can be classified into two categories: de novo and database search. By using dynamic programming, the de novo methods can predict peptide amino acid sequences directly from the spectral data.[7−10] This requires high-quality tandem mass spectra with high signal-to-noise ratios and mass accuracy. By contrast, the more prevalent database search programs predict peptide sequences by comparing the similarity between the observed spectra and virtual spectra generated from peptide sequences in a database. The predictions that database search programs make are influenced by the size, completeness, and redundancy status of the selected database. Sequest and Mascot are the most popular of these database search programs, are commercially available, and add peptide identification by mass spectrometry (MS) to the repertoire of many researchers.[11,12] Newer and alternative database search algorithms are available with purported advantages of providing improved probability-based scores, lower false-positive identification rates, enhanced speed, better modes for spectral interpretation, open source code, and so on.[13−16]

* To whom correspondence should be addressed. Phone: 301-504-9892. E-mail: cooperb@ba.ars.usda.gov.

[†] Johns Hopkins University.

[‡] USDA-ARS.

(1) Hunt, D. F.; Yates, J. R., III; Shabanowitz, J.; Winston, S.; Hauer, C. R. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 6233−6237.

(2) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., III. *Nat. Biotechnol.* **1999**, *17*, 676−682.

(3) Wolters, D. A.; Washburn, M. P.; Yates, J. R., III. *Anal. Chem.* **2001**, *73*, 5683−5690.

(4) Washburn, M. P.; Wolters, D.; Yates, J. R., III. *Nat. Biotechnol.* **2001**, *19*, 242−247.

(5) Evans, C. R.; Jorgenson, J. W. *Anal. Bioanal. Chem.* **2004**, *378*, 1952−1961.

(6) Hunt, D. F.; Buko, A. M.; Ballard, J. M.; Shabanowitz, J.; Giordani, A. B. *Biomed. Mass Spectrom.* **1981**, *8*, 397−408.

(7) Shevchenko, A.; Sunyaev, S.; Loboda, A.; Bork, P.; Ens, W.; Standing, K. G. *Anal. Chem.* **2001**, *73*, 1917−1926.

(8) Shevchenko, A.; Chernushevic, I.; Wilm, M.; Mann, M. *Mol. Biotechnol.* **2002**, *20*, 107−118.

(9) Mann, M.; Wilm, M. *Anal. Chem.* **1994**, *66*, 4390−4399.

(10) Shevchenko, A.; Chernushevich, I.; Wilm, M.; Mann, M. *Methods Mol. Biol.* **2000**, *146*, 1−16.

(11) Eng, J.; McCormack, A. L.; Yates, J. R., III. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−989.

(12) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551−3567.

10.1021/ac070202e CCC: $37.00 © 2007 American Chemical Society
Published on Web 04/19/2007

As a consequence of protein homology and record redundancy in databases, there are often many proteins that have nearly identical peptides, and database search results can be broadly interpreted to reflect the possibility that many of the proteins in the database were identified. However, there is a more realistic possibility that not all of the candidate proteins are present in an experimental sample. Thus, to balance what is possible and what is logical, it is necessary to report a nonredundant protein data set, whereby proteins having the same sets of matched peptides are grouped together and probabilities are calculated to assign confidence to assembled groups.[17−19] As a rule of thumb, grouped proteins are counted as a single protein and all proteins in a group are equal candidates for having been identified. To confidently identify any one candidate from another, experimental circumstances that preclude a candidate from being in the sample can be considered, or else additional distinguishing peptide information needs to be found.

To aid in the assembly of protein sets from peptides, algorithms have been developed to apply parsimony principles for groupings, and some algorithms apply probability values to evaluate the quality of the assemblies. For Sequest and Mascot users, DTASelect and DBParser, respectively, have been used to sort through and filter data sets.[20,21] While these programs reasonably apply parsimony principles to reduce redundancy, they do not provide probability-based information that can be used to assess the legitimacy of the assembled groups. ProteinProphet, on the other hand, has been used by both Sequest and Mascot users who seek to associate probability scores with protein groups.[22] As will be shown in this report, there are disadvantages to these software platforms as they tend to produce excessive false-positive data, waste acceptable true-positive data, or require an inordinate amount of computational effort. Thus, it remains a challenge for many researchers, especially Mascot users, to satisfy their own proteomics needs or meet publication standards for scientific journals.[19] In this light, we present a new probability model for protein assembly that factors peptide assignment scores provided by Mascot, the number of peptides in a database with similar molecular weights to a precursor ion, the number of times a sequence appears in a database, and the size of the search database. The algorithm then groups proteins having the same sets of matched peptides and calculates reasonable probabilities for grouped proteins with respect to the search database. The probabilities are in line with false-positive estimates that can be approximated by reverse database searching. The software platform, which we call PANORAMICS, can scale to size, and the data output is configured to provide desirable information in a simple format.

MATERIALS AND METHODS

Sample Preparation. Tryptic digests of bovine catalase, apotransferrin, carbonic anhydrase, glutamate dehydrogense, serum albumin, and lactoperoxidase were prepared according to manufacturer instructions (Michrom Bioresources, Auburn, CA). Approximately 10 pmol of each standard was evaluated. *Arabidopsis thaliana* plants were grown under standard growth conditions.[23] *Ustilago maydis* strain 521 was cultured in liquid potato dextrose broth. Plant leaves or collected fungal cells were frozen in liquid nitrogen and pulverized with a mortar and pestle. Proteins were precipitated in acetone/TCA, resolubilized in 8 M urea/100 mM Tris HCl pH 8.5, reduced, and carboxyamidomethylated.[24] An amount of 1 mg of soluble protein was digested with Lys-C (Roche Applied Science, Indianapolis, IN) and trypsin (Applied Biosystems, Foster City, CA), and peptides were desalted and concentrated by solid-phase extraction using SPEC-PLUS PT C18 columns (Varian, Lake Forest, CA).

Peptide Separation and Mass Spectrometry. Peptides were loaded onto pulled fused-silica capillaries that were packed with strong cation exchange and reversed-phase resins.[24] The column was placed in-line with a Surveyor HPLC pump, and peptides were eluted in a 12-step process that included increasing concentrations of salt followed by an increasing gradient of mobile phase at each step, as previously described.[24] The eluent was electrosprayed directly into the electrospray ionization source of the LCQ-Deca XP ion trap mass spectrometer. A parent ion scan was performed over the range of 400−1600 *m/z*. Automated peak recognition, dynamic exclusion, and MS/MS ion scanning of the top three most intense parent ions were performed using Xcalibur 1.2 (Thermo Electron, Waltham, MA), and tandem mass spectra were extracted from the raw data by Bioworks 3.1 (Thermo Electron). Parameters were set at the following: 400 minimum mass, 3500 maximum mass; 15 minimum ion count; 100 000 minimum TIC; 1.4 Da precursor mass tolerance; 1 group scan. All spectra not calculated as being singly charged were extracted as both doubly and triply charged spectra. A Perl script, merge.pl, which is part of the Mascot 2.1 software package (Matrix Science, London, U.K.), was used to convert multiple .dta files into a single file suitable for searching.

Database Searching. For the data described in the results, 2363 MS/MS spectra from the mixed protein standards, 35 875 spectra from *U. maydis*, 48 161 MS/MS spectra from yeast (kindly provided by Boris Zybailov and Mike Washburn at the Stowers Institute, Kansas City, MO, using an LCQ-Deca XP), and 15 775 from *A. thaliana* were compared with Mascot 2.1 to proteins in the 12/27/2005 release of the NCBI nonredundant (NR) protein database (3 148 822 sequences; ftp://ftp.ncbi.nih.gov/blast/db/), the *U. maydis* 521 NCBI RefSeq database (6528 sequences; http://www.ncbi.nlm.nih.gov/entrez/), the 11/26/2003 release of the *Saccharomyces cerevisiae* protein database (6479 sequences; ftp://ftp.ncbi.nih.gov/blast/db/), and/or the *A. thaliana* V. 6.0 protein database, herein called ATH1 (30 862 sequences; ftp://

[13] Zhang, N.; Aebersold, R.; Schwikowski, B. *Proteomics* **2002**, *2*, 1406−1412.
[14] Craig, R.; Beavis, R. C. *Bioinformatics* **2004**, *20*, 1466−1467.
[15] Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. *J. Proteome Res.* **2004**, *3*, 958−964.
[16] Tabb, D. L.; Narasimhan, C.; Strader, M. B.; Hettich, R. L. *Anal. Chem.* **2005**, *77*, 2464−2474.
[17] Nesvizhskii, A. I.; Aebersold, R. *Drug Discovery Today* **2004**, *9*, 173−181.
[18] Nesvizhskii, A. I.; Aebersold, R. *Mol. Cell. Proteomics* **2005**, *4*, 1419−1440.
[19] Carr, S.; Aebersold, R.; Baldwin, M.; Burlingame, A.; Clauser, K.; Nesvizhskii, A. *Mol. Cell. Proteomics* **2004**, *3*, 531−533.
[20] Tabb, D. L.; McDonald, W. H.; Yates, J. R., III. *J. Proteome Res.* **2002**, *1*, 21−26.
[21] Yang, X.; Dondeti, V.; Dezube, R.; Maynard, D. M.; Geer, L. Y.; Epstein, J.; Chen, X.; Markey, S. P.; Kowalak, J. A. *J. Proteome Res.* **2004**, *3*, 1002−1008.
[22] Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2003**, *75*, 4646−4658.
[23] Whitham, S. A.; Quan, S.; Chang, H. S.; Cooper, B.; Estes, B.; Zhu, T.; Wang, X.; Hou, Y. M. *Plant J.* **2003**, *33*, 271−283.
[24] Cooper, B.; Garrett, W. M.; Campbell, K. B. *Proteomics* **2006**, *6*, 2477−2484.

ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/). The yeast, *U. maydis*, and ATH1 databases were appended with sequences for common contaminant proteins such as human keratin. The amino acid sequences for each record in the databases were also reversed to create separate reverse sequence databases that were searched to approximate false identification rates.[25−27] One database contained the combined forward and reversed versions of the *A. thaliana* protein sequences. Searches were performed on a 7 node 3.2 GHz Dell server. Search parameters were set at fully tryptic digests and zero or one missed cleavages, and carboxyamidomethylation was selected as a fixed mass modification. Variable modifications such as oxidation of methionine were evaluated. Mass values were averaged, and peptide mass tolerance and fragment mass tolerance were set at ±1.5 and ±0.8 Da, respectively.

Assigning Probabilities to Peptides and Proteins. The major goal of our research was to design a model to determine the probability that proteins are present in a sample based upon the scores assigned to peptide sequence matches made by Mascot. Probability estimates of peptides and proteins are derived from these scores using the method outlined in the following three subsections. More details of the method, the algorithm, mathematical derivatives, and proofs are given in Supporting Information S1.

I. For each spectrum i, and for each peptide k in the database that is plausibly associated with it, the probability q_i^k that peptide k generates spectrum i is calculated.

Mascot .dat result files contain information for every tandem mass spectrum analyzed. The .dat file contains a list of qualified peptide sequences each with one or several scores measuring the similarity between the observed spectrum and the predicted spectrum of the peptide sequence. Mascot Ions scores are defined as follows,

$$S_i^k = -10 \log_{10}(P_i^k) \tag{1}$$

where P_i^k is the probability that the observed match of spectrum k to peptide sequence i is a random event, and where S_i^k is the Mascot Ions score for the peptide/spectrum match.[12] Despite this definition, we believe that P_i^k should be subject to more than just the similarity between the observed spectrum and the predicted spectrum. Other factors could include (but are not limited to) the scores of other peptide sequences corresponding to the same spectrum, the number of peptide sequences in the database that have a similar m/z ratio to the spectrum, the size of the database being searched, the charge state of the peptide, and the length of its sequence.

To make the peptide probabilities more accurate and theoretically more rigorous, we reformulated eq 1 and then redefined the probabilities of matches between peptide sequences and spectra. Equation 1 can also be written as,

$$\log_{10}(P_i^k) = -\frac{S_i^k}{10} \tag{2}$$

which shows that $\log_{10}(P_i^k)$ is a linear function of the Mascot Ions score S_i^k. We then generalize this linear function as follows,

$$\log(1 - q_i^k) = d_0 S_i^k + d_1 \log(T_i) + d_2 \log\log(M) + d_3 \tag{3}$$

where T_i is the number of peptides from the database that fall into the m/z ratio range of the precursor ion for tandem mass spectrum i and M is the total number of protein sequences in the selected database. Consideration of M in addition to T_i reduces effects of distraction and the acceleration of false-positive rates that occurs when large-sized databases are searched.[26,28] The log log (M) term is used to (roughly) keep a linear relationship with log T_i as database size increases. The constants d_0, d_1, and d_2 are used to balance the equation, and d_3 is the intercept value derived from linear regression analysis described in a later section. The variable q_i^k is the probability that peptide sequence k generated spectrum i, or in other words, that the match of peptide sequence k to the spectrum i is correct. To take into account the scores of other peptide sequences corresponding to the same spectrum, the probability q_i^k is further updated using an independent model (Supporting Information S1).

II. The q_i^k are used to calculate p_k, the probability that peptide k generated some spectrum.

When a peptide is matched to several spectra with the same charge state, the one with maximum probability is considered.[22] When a peptide exists in multiple charge states resulting from independent ionization events, the spectra with the maximum probabilities in all charge states are considered.[29] Using an independence assumption, this takes the value

$$p_k = 1 - \prod\nolimits_{\text{all charge states}} (1 - q_i^k) \tag{4}$$

where the product is over all spectra with different charge states matching to the peptide k. This equation is interpreted as saying that the probability that a peptide does not generate any spectra equals the probability that all matches of this peptide are false.

III. The probability π_m that some peptide in protein m generates some spectrum is solved.

By letting r_k^m denote the probability that peptide k from protein m generated some spectrum, an independence assumption gives

$$p_k = 1 - \prod\nolimits_m (1 - r_k^m) \tag{5}$$

and

$$\pi_m = 1 - \prod\nolimits_k (1 - r_k^m) \tag{6}$$

Equation 5 equates p_k, the probability that peptide k is matched to some spectrum correctly, to the probability that peptide k from

[25] Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. *J. Proteome Res.* **2003**, *2*, 43−50.

[26] Resing, K. A.; Meyer-Arendt, K.; Mendoza, A. M.; Aveline-Wolf, L. D.; Jonscher, K. R.; Pierce, K. G.; Old, W. M.; Cheung, H. T.; Russell, S.; Wattawa, J. L.; Goehle, G. R.; Knight, R. D.; Ahn, N. G. *Anal. Chem.* **2004**, *76*, 3556−3568.

[27] Elias, J. E.; Haas, W.; Faherty, B. K.; Gygi, S. P. *Nat. Methods* **2005**, *2*, 667−675.

[28] Cargile, B. J.; Bundy, J. L.; Stephenson, J. L., Jr. *J. Proteome Res.* **2004**, *3*, 1082−1085.

[29] Sonsmann, G.; Romer, A.; Schomburg, D. *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 47−58.

at least one protein generates some spectrum. Similarly, eq 6 equates π_m, the probability that some peptide in protein m generates at least one spectrum, to the probability that at least one peptide in the protein generates some spectrum.

If a sample contains two proteins m and n both of which contain shared peptide sequence k, under ideal experimental conditions the possibility of observing a spectrum acquired from peptide k only depends on the ionization and fragmentation conditions in the mass spectrometer and does not depend on whether the peptide originates from protein m or n. Using the definition of conditional probability, we then have the following relation:

$$\begin{aligned} r_k^m &= \text{prob}[\text{peptide } k \text{ from protein } m \\ &\qquad \text{generated some spectrum}] \\ &= \text{prob}[\text{peptide } k \text{ from protein } m \text{ generated some} \\ &\qquad \text{spectrum}|\text{protein } m \text{ generated some} \\ &\qquad \text{spectrum}]\text{prob}[\text{protein } m \text{ generated some spectrum}] \\ &= c_k \pi_m \end{aligned} \tag{7}$$

where c_k is a conditional probability the value of which only depends on peptide k. By replacing r_k^m with $c_k\ \pi_m$ in eqs 5 and 6, we obtain the following equation system,

$$\begin{aligned} p_k &= 1 - \prod\nolimits_m (1 - c_k \pi_m) \\ \pi_m &= 1 - \prod\nolimits_k (1 - c_k \pi_m) \end{aligned} \tag{8}$$

The system can be solved by an iterative algorithm to give expressions for π_m. The resulting solution for π_m is a reasonable probability that the protein m is present in the sample in relation to the selected database and the observed spectra.

Linear Regression Analysis. A separate set of 10 537 tandem mass spectra from six purified bovine protein standards was searched by Mascot against NCBI NR and the yeast database, ATH1, and a rice protein database (61 250 sequences, V3.0, http://rice.tigr.org/) appended with the six bovine protein records. This differential analysis was intended to account for Mascot Ions score differences with respect to different-sized databases to which correct and incorrect matches could be made. The following linear regression equation, which is a transformation of eq 3, was used to determine the parameters for the constants d_0, d_1, d_2, and d_3:

$$\log(1 - q_i^k) = d_0 \lambda (S_i^k - 10 \log T_i) + d_0 (1 - \lambda)(S_i^k - \chi) + d_3 \tag{9}$$

where $\chi = 47.8 \log \log M$ (the multiple 47.8 is the average of 10 log T_i from all databases divided by log log M of all databases). q_i^k was set at 0.9990 for spectra matching peptides from bovine standard sequences and 0 for nonbovine standard sequences (false-positives). No matches to peptide sequences from trypsin or contaminant proteins were used in linear regression. Because T_i and M are dependent variables, the sum of their respective constants was defined to be equal to the absolute value of d_0 to keep equivalent rates of change between database-dependent variables and Mascot Ions scores. Thus, $d_1 = 10d_0\lambda$ and $d_2 = 47.8d_0(1 - \lambda)$. The four parameters from the three-dimensional equation derived through the linear regression analysis are $d_0 = -0.3809$, $d_1 = 0.8858$, $d_2 = 13.97$, $d_3 = -7.458$. For all of our results presented here, our algorithm used this one set of fixed parameters. The following results demonstrate the effects of these parameters on a range of data sets searched against different-sized databases and for different organisms with underlying degrees of protein complexity.

Architecture of Software. PANORAMICS is the software package utilizing the probability algorithm. As web-based software, PANORAMICS has three major components all working as CGI programs. One is a file uploading program used to upload data files, such as Mascot .dat files. Other approaches to upload files such as ftp, scp, and Samba are also supported, which can allow for the transfer of files faster than the CGI program. The second component is the protein assembling program, the main program that runs on a chosen data file. The program output is an HTML file that allows for visualization of results using a web browser. The third component is the protein assembling revaluation program which accepts user input for known experimental conditions. All programs are written in ansi C programming language and compiled to run under Redhat Enterprise Linux AS 3.0 on a Dell PowerEdge 2800 server with two 3.4G CPUs, 2G RAM, and 360G hard disk space. Each program runs as a single process and a single thread. PANORAMICS supports multiple tasks, which permits several users to analyze different data sets independently at the same time. The number of tasks is only subject to limitations of the operating system and hardware. This software and a simplified version for Windows users is available directly from the authors.

RESULTS AND DISCUSSION

Analysis of Protein Standards. Mascot is a software platform that is used to interpret peptide tandem mass spectra.[12] Mascot compares observed spectra and virtual spectra derived from a database of protein sequences and provides a list of peptide sequences and Ions scores for these matches. When making peptide sequence/spectrum matches, Mascot considers precursor ion and daughter ion mass variation, amino acid mass variation, and enzyme cleavage patterns. A relationship between the numbers of peptides in the database with a similar precursor mass can be made to ascertain the quality of the assigned Ions scores. Although the details of this algorithm have never been published, the utility of Mascot has been described and it has generally fared well in comparison with other database search programs in the sense of making peptide sequence matches reliably.[27,30,31] In addition to making peptide sequence matches, Mascot can assemble the peptide sequences into candidate protein groups. We have not found this grouping function particularly useful for two reasons. For one, it operates from a CGI/Perl script application whose output is HTML. With large data sets, sometimes the size of the HTML output overwhelms the HTML browser. Two, it is often very apparent that proteins in an output list are redundant. For example, we analyzed an equal mixture of six known bovine protein standards by LC-MS/MS using tandem

(30) Chamrad, D. C.; Korting, G.; Stuhler, K.; Meyer, H. E.; Klose, J.; Bluggel, M. *Proteomics* **2004**, *4*, 619−628.

(31) Kapp, E. A.; Schutz, F.; Connolly, L. M.; Chakel, J. A.; Meza, J. E.; Miller, C. A.; Fenyo, D.; Eng, J. K.; Adkins, J. N.; Omenn, G. S.; Simpson, R. J. *Proteomics* **2005**, *5*, 3475−3490.

mass spectrometry methods previously described.[4,24] (This set of spectra was different from the set used for linear regression.) The data were searched by Mascot against the NCBI NR protein database. Mascot generated a list of 44 candidate proteins from the matched peptides (data not shown). Without going into the details of the list, it was very apparent that some candidate proteins were composed of peptide subsets belonging to another candidate protein. In other words, this list could possibly have been reduced to a smaller set of protein groups, but Mascot provided no easy way of accomplishing this. Other researchers have made similar observations and have made efforts to reduce the redundancy in Mascot protein lists.[20−22]

With these issues in mind, we designed a probability-based algorithm and software package called PANORAMICS whose input is the Mascot .dat results file generated from a database search and whose output is a list of nonredundant protein groups containing probability estimates that can be used to determine the likelihood that the protein assembly is correct with respect to the searched database. The use of PANORAMICS to analyze the .dat file for the previous six bovine protein standards resulted in a list of protein groups each associated with a calculated probability ranging from 1.00 to 0.00. The program output was configured to only show probabilities from 1.0000 to 0.8000, and there were 43 protein groups within this probability range (Supporting Information S2). Seventeen protein groups had probabilities ranging from 1.000 to 0.9500 (the latter being our imposed cutoff). The records for the 6 protein standards were included among the 17 protein groups. Bovine apotransferrin, lactoperoxidase, catalase, glutamate dehydrogenase, and albumin were found with probabilities of 1.0000, whereas carbonic anhydrase was found with a probability of 0.9999.

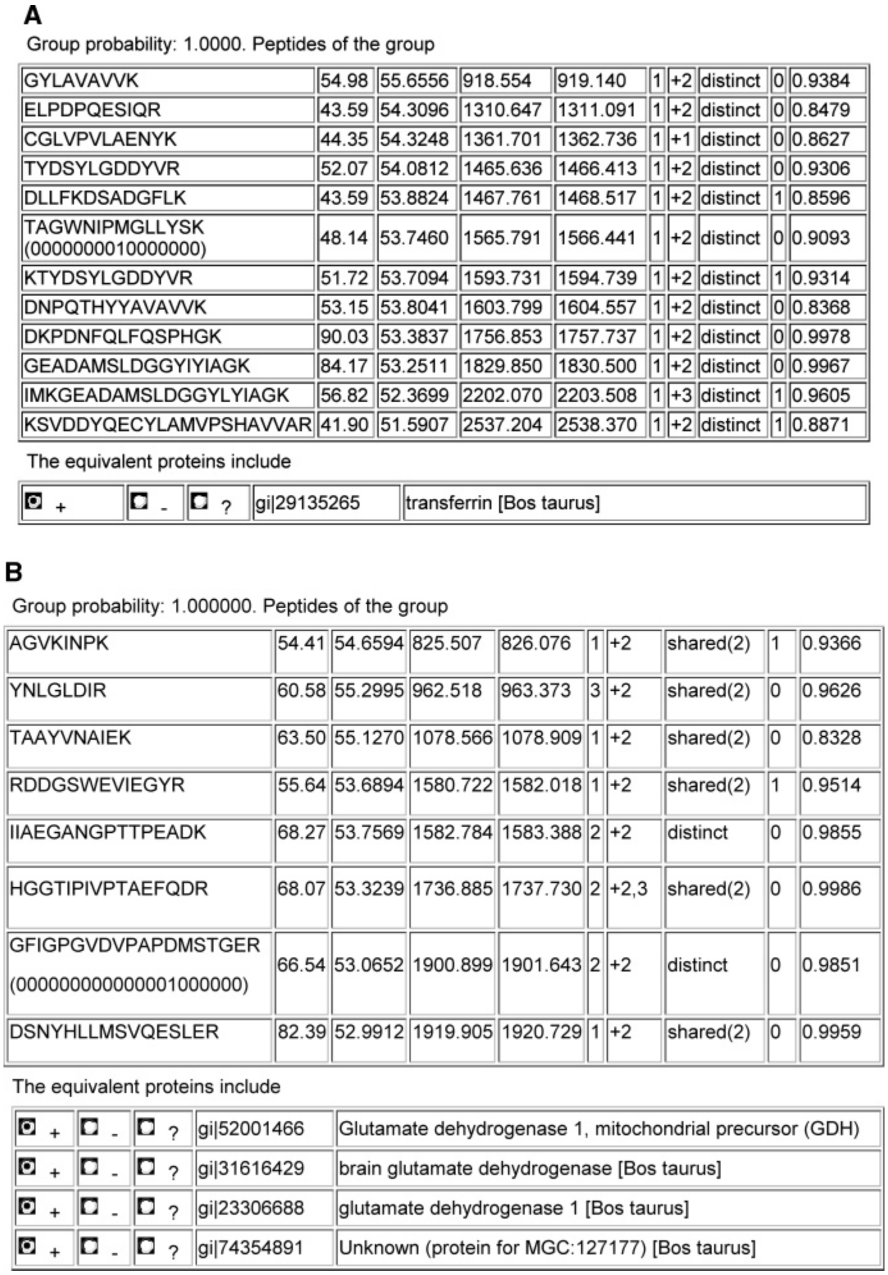
A closer look at PANORAMICS output from the protein standards data set shows the information used to qualify the groupings. For example, one group with probability 1.0000 comprised 12 peptides from the bovine transferrin protein gi|29135265 (Figure 1A). One of the peptides was interpreted by Mascot to have an oxidized methionine. The report also shows for each peptide the Mascot Ions score, Mascot Identity score, computed peptide mass, observed precursor mass, number of spectra assigned to the same sequence, the charge states that were found for spectra for each peptide, whether the matched peptide sequence is shared between protein groups or is distinct to a group, the number of missed cleavages in the sequence, and the probability of the particular peptide sequence. In this case, the group was assembled from the peptide sequences from a single record, and the large number of distinct peptides primarily contributed to the high probability score for this group.

Other factors besides the presence of distinct peptides can influence the protein group probability score. Generally, the detection of multiple charge states and high Mascot Ions scores with respect to the database leads to higher peptide probabilities. Higher peptide probabilities and more peptides can generally lead to greater group probabilities. Shared peptides also contribute to total protein group probabilities. But unlike distinct peptides, shared peptides do not contribute a full value of their peptide probability since this probability is distributed among the groups sharing it. The distribution of the shared probability is inherent to the protein probability algorithm, and the ultimate distribution is a result of algorithmic iteration over all protein groups assembled from the matched peptides. Even low scores can influence the protein group probability score. It has been general practice to discard matches with low scores, especially when peptide sequence assignments have an error rate of greater than 5% according to the Mascot Ions/Identity score relationship.[12,21,24] PANORAMICS, on the other hand, considers peptide matches with low scores since they have some probability, albeit a low one, of being correct.

If there is redundancy in a database, or if Mascot uses two or more separate protein records to interpret a spectrum, the data may not sufficiently distinguish these proteins. Therefore, proteins identified by the same set of matched peptides are grouped. For example, the protein standard glutamate dehydrogenase, gi|52001446, is one of the proteins that PANORAMICS grouped with other protein records contributing the same set of two distinct and six shared peptide sequences (Figure 1B). BLASTP analysis revealed that the other proteins in this group are 99%, 99%, and 97% identical at the amino acid level to record gi|52001446. Despite the minor sequence differences, there was not a sufficient amount of spectral data to distinguish one homologous glutamate dehydrogenase protein over another. Thus, this data can be interpreted to indicate that these proteins are indistinguishable and any one of them could be present in our sample. However, some of these proteins may not have existed in the sample. If we assume that only one was present in the standard sample (gi|52001446, according to the manufacturer), we can reasonably exclude the other possibilities from the group. Fortunately, PANORAMICS provides this option whereby the user can choose to keep or exclude a protein from the group. Doing so, however, only affects the list of proteins reported in a group and does not alter the group probability, unless all protein candidates are removed from a group, at which point PANORAMICS reassigns the shared peptides to other possible groups and recalculates probabilities. This beneficial feature is useful when many proteins are in a group and a user wishes to reduce the visual complexity of the output or if it is apparent from experimental circumstances that one protein is highly likely to be present compared to others.

In addition to the 6 protein standards, 11 other proteins were matched at a 95% or better confidence level. There are various explanations for the existence of these other proteins, the main one being that they were likely present as residual background proteins left over from the purification of the protein standards from blood or tissues. Most of these proteins match bovine records and could have easily been copurified. Two identified proteins, however, a camel peroxidase and human albumin, have 83% and 76% amino acid identity to the respective bovine standards. It is possible that variation in the collected bovine spectra resulted in Mascot making matches to the distinct peptide sequences from these protein records rather than the protein standards.[32] Or perhaps the exact bovine peptide records that best represent these spectra were not in the database, and the closest matches were to the human and camel records. One other important thing to note is that four of the protein groups were identified by single peptides, one of which was trypsin that was used to digest the standards. The PANORAMICS probability model allows single-peptide protein group identifications. But in order for these to

[32] Venable, J. D.; Yates, J. R., III. *Anal. Chem.* **2004**, *76*, 2928−2937.



A

Group probability: 1.0000. Peptides of the group

GYLAVAVVK	54.98	55.6556	918.554	919.140	1	+2	distinct	0	0.9384
ELPDPQESIQR	43.59	54.3096	1310.647	1311.091	1	+2	distinct	0	0.8479
CGLVPVLAENYK	44.35	54.3248	1361.701	1362.736	1	+1	distinct	0	0.8627
TYDSYLGDDYVR	52.07	54.0812	1465.636	1466.413	1	+2	distinct	0	0.9306
DLLFKDSADGFLK	43.59	53.8824	1467.761	1468.517	1	+2	distinct	1	0.8596
TAGWNIPMGLLYSK (0000000010000000)	48.14	53.7460	1565.791	1566.441	1	+2	distinct	0	0.9093
KTYDSYLGDDYVR	51.72	53.7094	1593.731	1594.739	1	+2	distinct	1	0.9314
DNPQTHYYAVAVVK	53.15	53.8041	1603.799	1604.557	1	+2	distinct	0	0.8368
DKPDNFQLFQSPHGK	90.03	53.3837	1756.853	1757.737	1	+2	distinct	0	0.9978
GEADAMSLDGGYIYIAGK	84.17	53.2511	1829.850	1830.500	1	+2	distinct	0	0.9967
IMKGEADAMSLDGGYLYIAGK	56.82	52.3699	2202.070	2203.508	1	+3	distinct	1	0.9605
KSVDDYQECYLAMVPSHAVVAR	41.90	51.5907	2537.204	2538.370	1	+2	distinct	1	0.8871

The equivalent proteins include

● +	○ -	○ ?	gi|29135265	transferrin [Bos taurus]

B

Group probability: 1.000000. Peptides of the group

AGVKINPK	54.41	54.6594	825.507	826.076	1	+2	shared(2)	1	0.9366
YNLGLDIR	60.58	55.2995	962.518	963.373	3	+2	shared(2)	0	0.9626
TAAYVNAIEK	63.50	55.1270	1078.566	1078.909	1	+2	shared(2)	0	0.8328
RDDGSWEVIEGYR	55.64	53.6894	1580.722	1582.018	1	+2	shared(2)	1	0.9514
IIAEGANGPTTPEADK	68.27	53.7569	1582.784	1583.388	2	+2	distinct	0	0.9855
HGGTIPIVPTAEFQDR	68.07	53.3239	1736.885	1737.730	2	+2,3	shared(2)	0	0.9986
GFIGPGVDVPAPDMSTGER (000000000000001000000)	66.54	53.0652	1900.899	1901.643	2	+2	distinct	0	0.9851
DSNYHLLMSVQESLER	82.39	52.9912	1919.905	1920.729	1	+2	shared(2)	0	0.9959

The equivalent proteins include

● +	○ -	○ ?	gi|52001466	Glutamate dehydrogenase 1, mitochondrial precursor (GDH)
● +	○ -	○ ?	gi|31616429	brain glutamate dehydrogenase [Bos taurus]
● +	○ -	○ ?	gi|23306688	glutamate dehydrogenase 1 [Bos taurus]
● +	○ -	○ ?	gi|74354891	Unknown (protein for MGC:127177) [Bos taurus]

Figure 1. Sample output from PANORAMICS. (A) Peptides identified by Mascot, all from a single bovine transferrin protein record (gi|29135265) from NCBI NR. The probability for this match is 1.0000. (B) Peptides identified by Mascot, all found in the same four records from NCBI NR with equal probability of 1.000000. The columns are arranged as follows: peptide sequence (along with a matching string of variable modifications, if applicable), Mascot Ions score, Mascot Identity score, computed peptide mass, observed precursor mass, number of tandem mass spectra assigned to the same peptide sequence, the charge states observed for the peptide, whether the peptide is shared between protein groups with different probabilities (number in parentheses is the number of groups peptides are shared with) or is distinct (considered a unique identifier for proteins grouped with the same probability), the number of missed tryptic cleavages in this sequence, and the probability for a particular peptide sequence. The first and last digits of the variable modification string correspond to the next N proximal and C proximal sequences outside of the listed sequence (i.e., eight numbers shown for six amino acids shown). Oxidized methionine was a variable modification for these searches. The user can click on the radio buttons for +/−/? to keep (with certainty (+) or uncertainty (?)) or remove proteins (−) from the data report.

receive high group probability rankings, the peptides must be assigned high Mascot Ions scores, the peptide sequence must be distinct, and the peptide sequence must not be too short in relation to the size of the database (the appearance of short sequences occurs more frequently as database size grows). All together, there is a sufficient amount of support information in the form of high Mascot scores, presence of distinct peptides, or multiple peptides found from each protein to justify the existence of these background and homologous proteins.

There is always the possibility that observed spectra are falsely associated with candidate peptide sequences,[12] but without manual inspection for each and every match, it can be difficult to determine false peptide/spectrum matches and subsequent false protein identifications. The probability estimates provided by PANORAMICS can be used to judge whether a protein identification is correct. To independently gauge the accuracy of the derived probability values, we used reverse database searching to approximate the rate of false peptide/spectrum matches and false protein groupings.[25–27] Any protein groups assembled by PANORAMICS from reverse database peptides would be false, unless their matching peptide sequences also appeared in the forward version of the database. In such a test, the number of groups found through reverse database searching that pass a 95% cutoff should not exceed 5% of the total number of groups found through forward database searching that exceed the same 95% threshold. When we searched the protein standards data set against the reversed NCBI NR database using Mascot and then processed the data through PANORAMICS, no protein was assigned a probability greater than 0.9500. The highest group was scored at 0.9365. These results support our hypothesis that PANORAMICS group probabilities are accurate measures of confidence for assessing the proteins assembled from Mascot peptide sequence inferences. This conclusion is also valid for the single-peptide protein groups whose protein probabilities are constrained to the sole peptide probabilities.

Analysis of Proteins from Fungi and a Model Plant. Yeast, *U. maydis*, and the model plant *A. thaliana* have different-sized sequenced genomes and differ in their cellular protein content and complexity. By comparing peptide tandem mass spectra from these organisms to small (yeast and *U. maydis*), midsized (ATH1), and very large-sized databases (NCBI NR), we can show the ability of PANORAMICS to logically assemble a probability-based set of candidate proteins from peptide data sets whose complex disorganization is compounded by the number of nonredundant protein entries in the database.

The yeast tandem mass spectra, produced by an outside lab, and the *U. maydis* spectra were analyzed by Mascot against the yeast and the *U. maydis* protein databases, respectively, and the results were compiled by PANORAMICS. For yeast, 573 protein groups were assembled at the 95% level while 559 were assembled for *U. maydis* (Supporting Information S3 and S4). For these data sets PANORAMICS accepted peptides as short as five amino acids; smaller peptide sequences would have a greater chance of appearing randomly in databases of these sizes. We performed the reverse database search, and PANORAMICS produced 35 protein groups at the 0.950 probability level or better for yeast and 25 for *U. maydis*. None of the groups had a probability of 1.0000; the highest were 0.9960 for yeast and 0.9937 for *U. maydis*.

Similarly, we tested Mascot data from a set of *A. thaliana* tandem mass spectra. Searches were made against specific forward and reversed versions of ATH1. PANORAMICS assembled 225 protein groups with better than 95% probability from the data set from the forward database search (Supporting Information S5, Table 1). Here, PANORAMICS accepted peptides as short as six amino acids; smaller peptide sequences would have a greater chance of appearing randomly in a database of this size. Reverse database searching and PANORAMICS analysis gave rise to three protein groups with a probability of 0.9500 or better (Supporting Information S6, Table 1). Again, none of the groups had a probability of 1.0000. As a separate analysis, we combined the ATH1 forward and reverse databases and searched against this single database that was now double in size. PANORAMICS assembled 218 protein groups with a probability in excess of 0.9500 (Supporting Information S7). Only one of those assembled protein groups contained peptides from the reversed sequences. Thus, there were six fewer (forward) protein groups at the 0.9500

Table 1. Analysis of Five Different Software Applications Used to Assemble Protein Groups from Mascot Database Search Results for Tandem Mass Spectra from *A. thaliana*[a]

	forward ATH1 protein database, protein groups	reverse ATH1 protein database, protein groups	false-positive rate estimate (%)[b]	run time (s)	forward NCBI NR protein database, protein groups	reverse NCBI NR protein database, protein groups	false-positive rate estimate (%)[b]	run time (s)
Mascot 2.1	361	46	12.7	60	1082	41	3.79	1320[c]
DTASelect 1.9 1 peptide/protein peptide $p < 0.05$	310	39	12.6	1	300	35	11.6	50
DTASelect 1.9 2 peptide/protein peptide $p < 0.05$	124	0	0	1	162	0	0	45
DBParser 2.0 1 peptide/protein peptide $p < 0.05$	348	40	11.5	480	not performed	not performed		terminated by users after 22 h
DBParser 2.0 2 peptide/protein peptide $p < 0.05$	88	0	0	480	not performed	not performed		terminated by users after 22 h
ProteinProphet Protein 95%	228[d]	0[d]	0	2067	138[d]	0[d]	0	4570
PANORAMICS Protein 95%	225[d]	3[d]	1.3	1	202[d]	5[d]	2.5	6

[a] Calculated false-positive rates and software application run times are shown. [b] Measured by dividing the number of groups of proteins from a reverse database search by the number of groups of proteins from a forward database search. [c] As computed by master_results.pl, which is part of the Mascot software package. [d] Groups with a probability equal to or greater than 95%.

level in the combined forward and reverse database search than in the forward ATH1 search. However, these six protein groups still remained in the whole PANORAMICS data set; they just received lower probabilities because of adjustments made by PANORAMICS to compensate for the larger database. For example, protein group no. 225 with probability 0.9507 was listed as protein group no. 234 with probability 0.9278 in the data set from the combined forward and reverse database search.

Continuing with our analysis on the effects of database size on search results, we used Mascot to search the *A. thaliana* spectra against the forward and reverse NCBI NR databases and analyzed the data with PANORAMICS. At the group probability level of greater than 0.9500, 202 protein groups were assembled from the forward database search and 5 protein groups were assembled from the reverse database search (Supporting Information S8 and S9, Table 1). Thus, by comparison to the assemblies made from searches against the much smaller ATH1 forward database, there were 23 fewer protein groups at the 0.950 probability level. Such a finding is not unusual, however. NCBI NR contains not only the same records as ATH1 but also records from other plants and organisms to which matches can be made. Consequently, the NCBI NR database is 100 times larger than ATH1, and as the size of databases grows, the chance for Mascot to generate random matches becomes larger. To counteract this effect, PANORAMICS considers database size to control false-positive rates. As a result, the minimum peptide length in the PANORAMICS report changed from six to seven amino acids, which in turn affected the protein group assembly. Hence, it is difficult to compare the assembled protein groups from the NCBI NR and ATH1 searches because the variation between the data sets is a result of Mascot searching against very different databases tempered by PANORAMICS adjustments for the database size. Because of this, it is probably best to search against a database that suits an application and then allow PANORAMICS to limit false-positive identifications with respect to database size.

Other Protein Assembly Software. We have compared the output from several software applications to PANORAMICS using the data from the Mascot searches of the *A. thaliana* spectra against the forward and reverse ATH1 and NCBI NR databases. The findings are summarized in Table 1 and discussed below.

Comparison to Mascot Protein Assembly. The internal Mascot protein assembly function runs off a CGI/Perl script (master_results.pl) that is part of the Mascot software package. It produced 361 proteins from the forward search against ATH1 and 46 proteins in the corresponding reverse database search (Supporting Information S10 and S11, Table 1). This represents a 12.7% protein false-positive rate. When analyzing the very large .dat file for the NCBI NR searches, the HTML output from master_results.pl overwhelmed our web browser. By manually interpreting the file we determined that there were 1082 proteins from the forward NCBI NR search and 41 proteins from the reverse search (Supporting Information S12 and S13, Table 1). Although this translates to a 3.79% false-positive rate at the protein level, this false-positive estimate is misleading because the number of proteins in the assembly from the forward search is higher than it could be. Since NCBI NR contains the same records as ATH1 plus others, additional protein assemblies arose due to the presence of homologous sequences in the database. Had parsimony been rigorously applied or if the shared proteins had not been equally distributed among all proteins harboring those sequences, the number of assembled proteins would have been much lower and the false-positive rate would have risen proportionately.

Comparison to DTASelect and DBParser Protein Assembly. DTASelect V. 1.9,[20] a widely used and useful program for sorting Sequest output, has partial functionality with Mascot data if the output file is first converted using a data export feature found in Mascot V. 2.1. Using the parsimony option and the parameters that all Mascot peptide sequence inferences must exceed the 0.05 false match cutoff established by the Mascot Ions/Identity score relationship ($p \leq 0.05$) and that there must be at least one peptide per group, DTASelect assembled 310 protein groups from the forward ATH1 search and assembled 39 protein groups from the reverse ATH1 search (Supporting Information S14 and S15, Table 1). This coincides with a 12.6% false identification rate at the protein level. DTASelect was able to sort the data in a second, but it took Mascot 4 min to configure the .dat file for DTASelect searching. DTASelect also assembled 300 proteins from the forward NCBI search data and 35 proteins from the reverse NCBI search data, which corresponded with an 11.6% false-positive rate (Supporting Information S16 and S17, Table 1). It took Mascot 20 min to configure the .dat file and 50 s for DTASelect to operate.

We also tested DBParser V. 2.0,[21] which functions similarly to DTASelect. Assembling proteins from peptides with Mascot probability scores below the 0.05 random match threshold (established by the Mascot Ions/Identity score relationship; $p \leq 0.05$), DBParser recognized 348 nonredundant protein groups from the forward ATH1 search (Supporting Information S18, Table 1). By contrast, 40 nonredundant proteins were found from the reverse search, and this is equivalent to an 11.5% false-positive rate at the protein level (Supporting Information S19, Table 1). It took 8 min for DBParser to parse the data and generate the report for the forward data set and a little less time for the reverse data set. We attempted to analyze the Mascot report from the searches against the NCBI NR database, but after 22 h of DBParser processing and no indication of when the task would finish, we aborted the effort (Table 1). We examined the code for the DBParser algorithm and attribute the slow processing time to inefficient communication between Perl and the MySQL database system used to store data records.

When using protein assembly programs like DTASelect and DBParser, proteins most likely to be false-positive are those identified by only one peptide.[33] However, since DTASelect and DBParser work as filters, parameters can be changed to improve protein assembly confidence. By requiring two peptides for any protein identification, the false-positive rates were lowered to zero. (Supporting Information S20, S21, and S18 and S19, Table 1). This was not surprising since the chance of assembling a protein from two falsely identified peptides is very small. However, many protein assemblies were sacrificed in the process compared to PANORAMICS. Again, we were not able to judge DBParser against NCBI NR data for the reasons stated previously.

[33] Veenstra, T. D.; Conrads, T. P.; Issaq, H. J. *Electrophoresis* **2004**, *25*, 1278–1279.

Comparing the DTASelect and DBParser results to PANORAMICS results is not straightforward because PANORAMICS considers the probability that shared peptides are more likely to belong to one protein over another. DTASelect and DBParser analyses do not guarantee this possibility. Rather they assume all peptides passing a threshold have equal probability. Also, while there is no confidence score generated for single-peptide assemblies from DTASelect and DBParser, PANORAMICS balances the acquisition of false-positive and valid data. Specifically, PANORAMICS will provide high probability scores to proteins identified by single peptides when (i) Mascot Ions scores are sufficiently high in comparison to the size of the database searched, (ii) when the number of peptides in the database with similar molecular weights is sufficiently small, and (iii) when the peptide sequence is distinct to a protein group. Consequently, assembled PANORAMICS protein groups can be remarkably different compared to those from DTASelect and DBParser. However, PANORAMICS users can have data that is controlled as stringently or as leniently as a filtering program's simply by choosing an appropriate probability level for acceptance.

Comparison to ProteinProphet Protein Assembly. The software package of PeptideProphet/ProteinProphet is also a popular tool for evaluating Mascot and Sequest output.[22,34] We installed it on a desktop computer in May, 2006 according to instructions provided by the Institute of Systems Biology (Seattle, WA). This software employs a statistical model and provides probabilities as confidence measures for peptide and protein identifications. ProteinProphet reported 228 protein groups beyond the 0.95 probability threshold for the forward ATH1 database search and 138 nonredundant protein groups for the NCBI NR forward database search (Supporting Information S22 and S23, Table 1). ProteinProphet did not report any protein groups with probabilities greater than 0.95 for searches against either corresponding reverse database (Supporting Information S24 and S25, Table 1). This points to an inaccuracy with the ProteinProphet probability model since probability dictates there should be some false-positives at this threshold.

To examine the protein assemblies between the two models more closely, we compared all proteins with probabilities ranging from 1.0 to 0.20 computed from the forward searches to those from the reverse searches (Figure 2). For both data sets searched against ATH1 and NCBI NR, ProteinProphet generated as many or fewer false-positives when proteins were specifically identified at a probability level close to 1.0 but produced more false-positive and less true-positive data overall. For example, when analyzing the *A. thaliana* data set searched against NCBI NR, ProteinProphet identified 262 proteins at a 0.60 probability cutoff but 353 proteins from the reverse database (Figure 2B). A larger number of protein identifications from the reverse database search indicates low selectivity for ProteinProphet. This phenomenon combined with the sharp turns in the ROC plots shows that the probabilities produced by ProteinProphet are not accurate for these analyses. By contrast PANORAMICS identified 540 proteins above a 0.60 probability level and 204 from the reverse database search (Figure 2B). Thus, PANORAMICS is much more accurate in terms of selectivity than ProteinProphet.

At a high probability level, however, ProteinProphet appeared to be more sensitive. But we believe that this is an erroneous attribute. In both PANORAMICS' and Protein Prophet's probability model, if protein A has two matched peptides that are distinct, then the probability that protein A is present is computed by

$$\begin{aligned}
&\mathrm{P}(\text{protein A is present})\\
&= \mathrm{P}(\text{peptide 1 is matched correctly OR}\\
&\qquad\qquad \text{peptide 2 is matched correctly})\\
&= 1 - \mathrm{P}(\text{neither peptide 1 nor}\\
&\qquad\qquad \text{peptide 2 is matched correctly})\\
&= 1 - \mathrm{P}(\text{peptide 1 is a false match and}\\
&\qquad\qquad \text{peptide 2 is a false match})\\
&= 1 - \mathrm{P}(\text{peptide 1 is a false match})\mathrm{P}\\
&\qquad\qquad (\text{peptide 2 is a false match})\\
&= 1 - (1 - \mathrm{P}(\text{peptide 1 is a true match}))\\
&\qquad\qquad (1 - \mathrm{P}(\text{peptide 2 is a true match})) \quad (10)
\end{aligned}$$

which is denoted by

$$\mathrm{P}(\text{protein A is present}) = 1 - (1 - p_1)(1 - p_2) \quad (11)$$

However, ProteinProphet then updates the peptide probabilities by considering an estimated number of sibling peptides (NSP) and then calculates a protein probability using an independent model,[22] while PANORAMICS does not use any information from sibling peptides to update probabilities. "Updating" peptide probabilities by considering NSPs changes the meaning of the probabilities p_1 and p_2. In this case, p_1, which is the unconditional probability that peptide 1 is a true match, turns into a conditional probability p_1^*, the probability that peptide 1 is a true match knowing the fact that peptide 2 is matched. The updated p_1^* is

$$p_1^* = \mathrm{P}(\text{peptide 1 is a true match}|\text{peptide 2 is matched}) \quad (12)$$

The same thing happens to p_2^*. Although p_1^* and p_2^* are better probability estimates for peptides 1 and 2 given all the assignment information, the change from unconditional to conditional probability precludes them for use in eq 11. Thus, using NSP gives more benefit to protein groups with multiple peptides. Consequently, it gives fewer false-positives when a very high probability threshold is set, but the accuracy of the probability values given by ProteinProphet deteriorates (Figure 2B).

Thus, it is apparent that PANORAMICS can assemble more protein groups from a data set searched against NCBI NR than ProteinProphet can and that the basis for this difference may be due to ProteinProphet's flawed sensitivity. However, the ProteinProphet data set is not a complete subset of the PANORAMICS data set, and a closer examination of the specific assemblies themselves reveals the stark differences between ProteinProphet and PANORAMICS scoring. Using the *A. thaliana* spectra searched against NCBI NR as an example, there were 24 proteins unique to the ProteinProphet data set that did not pass the PANORAMICS 0.9500 probability threshold. For example, Pro-

[34] Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383−5392.

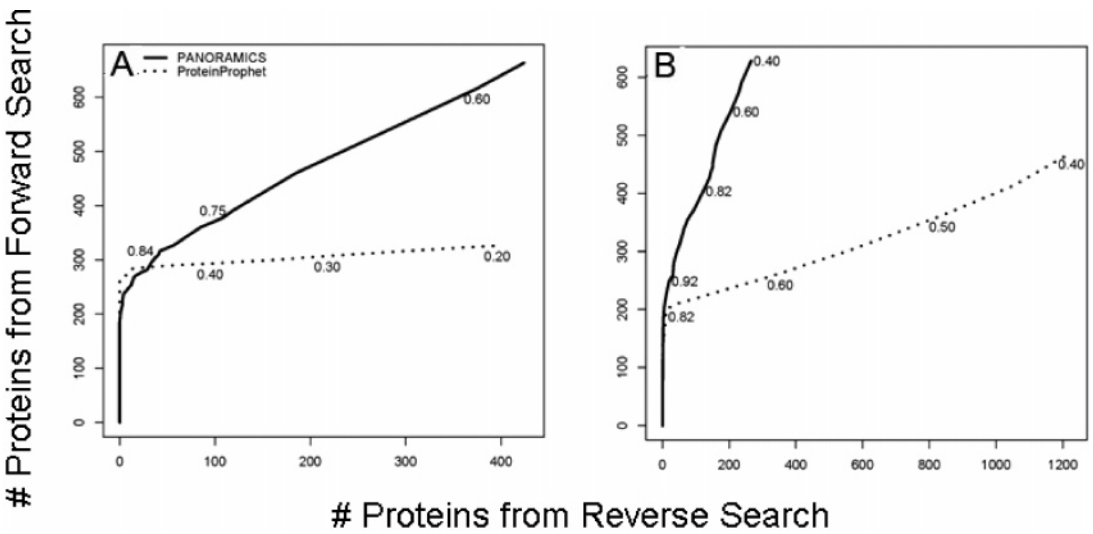


Figure 2. Comparison between the number of proteins identified from forward database searches vs reverse database searches for PANORAMICS and ProteinProphet. The same spectral data set from *A. thaliana* was used for all searches. Every point in the curves represents the number of proteins identified in a forward database search (*y*-axis) and the number of proteins identified in reverse database search (*x*-axis) corresponding to the same probability threshold. Some of the selected probability threshold values are shown. (A) ATH1 database search. (B) NCBI NR database search.

teinProphet protein group no. 99 with probability of 0.99 is protein group no. 306 with probability 0.8857 in PANORAMICS (Supporting Information S23 compared to Supporting Information S8). Whereas ProteinProphet assigned 0.96 and 0.78 probabilities to the two peptides contributing to this group's assembly, PANORAMICS assigned a probability of 0.8857 to the highest scoring short peptide and did not consider the second peptide because its Mascot Ions score was too low with respect to the size of the NCBI NR database. In another example, ProteinProphet protein group no. 9 with probability 1.00 was not found in the top 90% of groups assembled by PANORAMICS (Supporting Information S23 compared to Supporting Information S8). The ProteinProphet group no. 9 was assembled from the same two peptide sequences that differ by the identification of a variable oxidation modification at methionine. While ProteinProphet counted these two peptides separately to obtain a high group probability, PANORAMICS counted these two peptides varying only in a post-translational modification as one, thus resulting in a lower group probability score. As a final example, ProteinProphet group no. 10 received a high probability because the sequence AYGXAANVFGKPK was considered distinct despite the fact that there was an ambiguous amino acid in the sequence. By contrast, PANORAMICS disregarded this sequence because the classification for this peptide is actually unresolved, even though the sequence string itself is unique to the database.

Again, some differences in the two algorithms' scoring mechanisms may account for some of these observed discrepancies. PeptideProphet derives a discriminant score for each peptide assignment from the several scores provided by the database search program. By assuming the discriminant scores are normal-distributed random variables for true matches and Γ-distributed for false matches, PeptideProphet uses the expectation maximization (EM) algorithm to estimate the probability distributions' parameters.[22,34] However, by only considering the discriminant scores of peptides, this procedure may break the relationship between a spectrum and the peptides assigned to it. For PANORAMICS, this relationship works as a constraint for the probabilities of the peptides corresponding to a single spectrum, especially when a spectrum is matched by a database search program to several sequences with high scores. Another difference lies in the probability model of ProteinProphet, which assumes that if two or more protein groups share a peptide sequence, only one of them can be present in the sample. Our lab experience suggests that a peptide sequence could be found in many heterologous proteins. In such cases, it might be more appropriate to say that this type of peptide is quite universal and cannot be used to distinguish different proteins.

In terms of computing time for the analysis of forward NCBI database search results, it took 39 min to transform the Mascot .dat to an .xml file, 7 min to run PeptideProphet, and 31 min to run ProteinProphet, or 76 min in total. By contrast, it took PANORAMICS 6 s to run the whole procedure on the same computer (Table 1).

CONCLUSIONS

Tandem mass spectrometry coupled with HPLC separation usually generates a great amount of spectral data that can be used to identify proteins in a mixture. The development of efficient algorithms that provide proper confidence measures for protein identifications is essential to this process. The algorithm described here is based on a probability model that calculates probability measures for the protein identifications using peptide sequence inference scores generated by a popular database search engine. The probability values provide a simple and standard measure for the estimation of false-positive rates. The algorithm specifically adjusts for database size, thereby allowing researchers to perform proteomics experiments on nonmodel organisms requiring large databases for the interpretation of their spectra. The software using this algorithm, PANORAMICS, was also specifically designed to operate quickly and scale to size, which brings more convenience to and extends the frontier of using tandem mass spectrometry for the analysis of more complex and larger-scaled systems.

The mass similarities among the 20 amino acids challenge all database search programs and de novo programs used for spectral interpretation. Isoleucine cannot be readily differentiated from leucine, and some amino acid pairs (I, N and Q, K) cannot be distinguished if the mass resolution is larger than 1 Da. This creates a circumstance where two distinct peptides can match the same spectra. To solve the problem of generating excessive false-positive data by allowing different peptide sequences to match the same spectra, the PANORAMICS algorithm employs new concepts for peptide grouping. Two peptide sequences are said to be in the same peptide group if their predicted spectra are not distinguishable. Concepts for peptide grouping also carry through to the protein grouping level, where two proteins are said to be in the same protein group if their matched peptide sets are the same.[21,22] Proteins that fall into the same group are often homologous proteins which cannot be distinguished without additional mass spectral information. Applying these concepts of peptide and protein groups together reduces false-positive rates and reduces an unwanted side effect of having redundant protein information. As a result, PANORAMICS works efficiently for data from searches against nonredundant or redundant databases. In the end, parsimony is achieved when using this program.

The model we present integrates peptide probabilities with protein probabilities. In our model, even peptides with low probabilities are considered in the protein assemblies. This is in direct contrast to popular filtering methods whereby peptides with probabilities below a threshold are excluded from a protein assembly. One of our conclusions drawn from evaluating protein assembly programs using a peptide probability filter is that protein groups assembled from peptides exceeding a probability threshold do not always inherit a probability exceeding the threshold. This implies that rigorous attention applied to accurately measuring peptide/spectrum matches matters little if the results are not taken within the context of identifying proteins, if identifying proteins is the aim of an experiment.

To gauge the accuracy of our model, we have performed reverse database searching as an independent measure for approximating false-positive rates. The reverse database search has been used for modeling because the reverse database has the same distribution of the number of amino acids and the same number of nonsensical protein records. But the final fact is that this type of simple analysis is not completely suited to measure the accuracy of an integrated peptide/protein probability model such as ours. Reverse database searching tends to lead to an overestimation of the false-positive rate implied by PANORAMICS probabilities. Under the normal circumstances, the "correct" match between a spectrum and a peptide in a database usually produces the highest (or almost highest) score, making this match easily distinguishable from the random "incorrect" matches. On the basis of this fact, PANORAMICS uses the higher scores for peptides matched to the same spectrum as a "firewall" to prevent false-positives. However, all matches in reverse database searches, excluding any true-positives, are false matches. This fact makes the "firewall" effect in PANORAMICS disappear, and the false-positive rate goes up as a result. Due to this reasoning, a better, although not perfect, way to measure the false-positive rate is to search a combined forward and reverse database. We show that this method can give a lower value for false-positive rate estimates than the normal reverse database search. However, concatenating databases increases the size of the database, causing PANORAMICS to overcompensate in reducing false-positive rates. Nonetheless, we contend that it is not necessary to routinely perform reverse database searches or combined forward/reverse database searches to estimate false-positive rates because PANORAMICS provides reasonable probability scores that can in turn be used to evaluate data quality.

We expect our chosen parameters to be suitable for a variety of sample conditions and data set sizes, as demonstrated. Release of programming code is planned for the future. Finer adjustments can be made by users, but should be independently evaluated, especially for applications and instruments that drastically deviate from what we have described.[35] The described version of PANORAMICS runs on Linux or Windows XP operating systems and supports Mascot V. 2.0, 2.1, and 2.2 .dat output files and is freely available directly from the authors.

ACKNOWLEDGMENT

We thank Dr. Alan J. Goldman from the Johns Hopkins University and Drs. Neerav Padliya and Joohyun Lee at USDA-ARS, Beltsville, MD for helpful discussions and preparation of this manuscript. We also thank Drs. Boris Zybailov and Mike Washburn at the Stowers Institute, Kansas City, MO for yeast tandem mass spectra. This work was funded by a specific cooperative agreement between the Johns Hopkins University and USDA-ARS, Beltsville, MD.

SUPPORTING INFORMATION AVAILABLE

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review January 31, 2007. Accepted March 22, 2007.

AC070202E

[35] Xie, H.; Griffin, T. J. *J. Proteome Res.* **2006**, *5*, 1003−1009.