Anal. Chem. **2003,** *75,* 4646−4658

A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry

Alexey I. Nesvizhskii,*,† Andrew Keller,*,† Eugene Kolker,‡ and Ruedi Aebersold

Institute for Systems Biology, 1441 North 34th Street, Seattle, Washington 98103

A statistical model is presented for computing probabilities that proteins are present in a sample on the basis of peptides assigned to tandem mass (MS/MS) spectra acquired from a proteolytic digest of the sample. Peptides that correspond to more than a single protein in the sequence database are apportioned among all corresponding proteins, and a minimal protein list sufficient to account for the observed peptide assignments is derived using the expectation−maximization algorithm. Using peptide assignments to spectra generated from a sample of 18 purified proteins, as well as complex *H. influenzae* and *Halobacterium* samples, the model is shown to produce probabilities that are accurate and have high power to discriminate correct from incorrect protein identifications. This method allows filtering of large-scale proteomics data sets with predictable sensitivity and false positive identification error rates. Fast, consistent, and transparent, it provides a standard for publishing large-scale protein identification data sets in the literature and for comparing the results obtained from different experiments.

The goal of proteomics is to identify and characterize all proteins expressed in cells grown under a variety of conditions.[1] Tandem mass spectrometry (MS/MS) has become the method of choice for identification of proteins in high-throughput proteomics studies.[2,3] It has been particularly useful for cataloging and quantifying proteins in a number of organisms,[4−7] for protein complex characterization and protein−protein network reconstruction,[8−11] for studying signaling pathways,[12] and for metabolic pathway reconstruction.[13,14] Finally, mass spectrometry is expected to play an important role in the ambitious task of modeling cell behavior.[15]

A general view of the MS/MS-based approach to study complex protein mixtures is illustrated in Figure 1. Sample proteins are first proteolytically cleaved into smaller peptides, most often by the enzyme trypsin. Protein digestion is required because intact proteins are not amenable for mass spectrometric identification, though some progress toward removing this limitation has been recently reported.[16,17] Complexity of the peptide mixture can be reduced by strong cation exchange chromatography or other available separation techniques. The resulting peptide mixture is then subjected to reversed-phase chromatography directly coupled with a mass spectrometer. Alternatively, peptides eluting from the reversed-phase column can be deposited on a plate for subsequent analysis by MALDI-MS/MS. Peptides are then ionized and selected ions subjected to fragmentation in the collision cell to produce tandem mass spectra. At this stage, computational methods must be used to infer the peptides and proteins that gave rise to the observed spectra.[18,19] Database search programs such

* Corresponding authors: (e-mail) nesvi@systemsbiology.org; akeller@systemsbiology.org.(fax) 206-732-1299.

† Contributed equally to this work.

‡ Current address: BIATECH, North Creek Parkway, Suite 115, Bothell, WA 98011.

(1) Pandey, A.; Mann, M. *Nature* **2000**, *405*, 837−846.

(2) Aebersold, R.; Goodlett, D. R. *Chem. Rev.* **2001**, *101*, 269−287.

(3) Smith, R. D.; Anderson, G. A.; Lipton, M. S.; Masselon, C.; Paša-Tolić, L.; Shen, Y.; Udseth, H. R. *OMICS* **2002**, *6*, 61−90.

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

(5) Lipton, M. S.; Paša-Tolić, L.; Anderson, G. A.; Anderson, D. J.; Auberry, D. L.; Battista, J. R.; Daly, M. J.; Fredrickson, J.; Hixson, K. K.; Kostandarithes, H.; Masselon, C.; Markillie, L. M.; Moore, R. J.; Romine, M. F.; Shen, Y.; Stritmatter, E.; Tolić, N.; Udseth, H. R.; Venkateswaran, A.; Wong, K. K.; Zhao, R.; Smith, R. D. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 11049−11054.

(6) Lasonder, E.; Ishihama, Y.; Andersen, J. S.; Vermunt, A. M. W.; Pain, A.; Sauerwein, R. W.; Eling, W. M.; Hall, N.; Waters, A. P.; Stunnenberg, H. G.; Mann, M. *Nature* **2002**, *419*, 537−542.

(7) Florens, L.; Washburn, M. P.; Raine, J. D.; Anthony, R. M.; Grainger, M.; Haynes, J. D.; Moch, J. K.; Muster, N.; Sacci, J. B.; Tabb, D. L.; Witney, A. A.; Wolters, D.; Wu, Y.; Gardner, M. J.; Holder, A. A.; Sinden, R. E.; Yates, J. R.; Carucci, D. J. *Nature* **2002**, *419*, 520−526.

(8) 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.

(9) Gavin, A.; Bosche, M.; Krause, R.; Grandi, P.; Marzioch, M.; Bauer, A.; Schultz, J.; Rick, J. M.; Michon, A. M.; Cruciat, C. M.; Remor, M.; Hofert, C.; Schelder, M.; Brajenovic, M.; Ruffner, H.; Merino, A.; Klein, K.; Hudak, M.; Dickson, D.; Rudi, T.; Gnau, V.; Bauch, A.; Bastuck, S.; Huhse, B.; Leutwein, C.; Heurtier, M. A.; Copley, R. R.; Edelmann, A.; Querfurth, E.; Rybin, V.; Drewes, G.; Raida, M.; Bouwmeester, T.; Bork, P.; Seraphin, B.; Kuster, B.; Neubauer, G.; Superti-Furga, G. *Nature* **2002**, *415*, 141−147.

(10) Ho, Y.; Gruhler, A.; Heilbut, A.; Bader, G. D.; et al. *Nature* **2002**, *415*, 180−183.

(11) Ranish, J. A.; Yi, E. C.; Leslie, D. M.; Purvine, S. O.; Goodlett, D. R.; Eng, J.; Aebersold, R. *Nat. Genet.* **2003**, *33*, 349−355.

(12) Pandey, A.; Podtelejnikov, A. V.; Blagoev, B.; Bustelo, X. R.; Mann, M.; Lodish, H. F. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 179−184.

(13) Baliga, N. S.; Pan, M.; Goo, Y. A.; Yi, E. C.; Goodlett, D. R.; Dimitrov, K.; Shannon, P.; Aebersold, R.; Ng, W. V.; Hood, L. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 14913−14918.

(14) Guina, T.; Purvine, S. O.; Yi, E. C.; Eng, J.; Goodlett, D. R.; Aebersold, R.; Miller, S. I. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 2771−2776.

(15) Forst, C. V. *Mol. Biol. Rep.* **2002**, *29*, 265−280.

(16) Kelleher, N. L.; Lin, H. Y.; Valaskovich, G. A.; Aaseruud, D. J.; Fridriksson, E. K.; McLafferty, F. W. *J. Am. Chem. Soc.* **1999**, *121*, 806−812.

(17) Meng, F.; Cargile, B. J.; Patrie, S. M.; Johnson, J. R.; McLoughlin, S. M.; Kelleher, N. L. *Anal. Chem.* **2002**, *74*, 2923−2929.

(18) Fenyo, D. *Curr. Opin. Biotechnol.* **2000**, *11*, 391−395.

(19) Chakravarti, D. B.; Chakravarti, B.; Moutsatsos, I. *Biotechniques* **2002**, (Suppl. 32), 4−15.

10.1021/ac0341261 CCC: $25.00 © 2003 American Chemical Society
Published on Web 07/15/2003

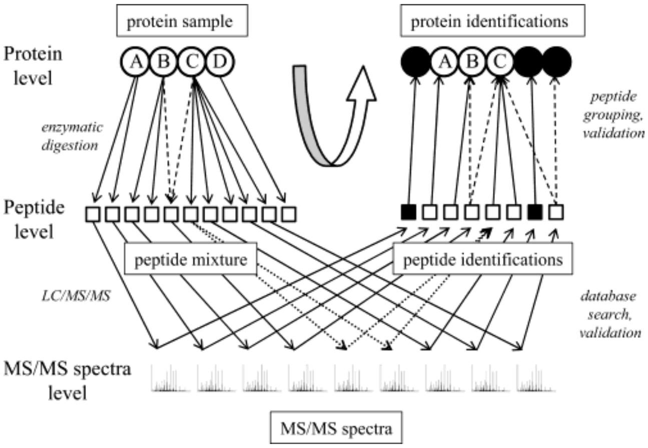


Figure 1. Simplified outline of the experimental steps and flow of the data in a typical high-throughput mass spectrometry-based analysis of complex protein mixtures. Each sample protein (open circle) is cleaved into smaller peptides (open squares), which can be unique to that protein or shared with other sample proteins (indicated by dashed arrows). Peptides are then ionized and selected ions fragmented to produce MS/MS spectra. Some peptides are selected for fragmentation multiple times (dotted arrows) while some are not selected even once. Each acquired MS/MS spectrum is searched against a sequence database and assigned a best matching peptide, which may be correct (open square) or incorrect (black square). Database search results are then manually or statistically validated. The list of identified peptides is used to infer which proteins are present in the original sample (open circles) and which are false identifications (black circles) corresponding to incorrect peptide assignments. The process of inferring protein identities is complicated by the presence of degenerate peptides corresponding to more than a single entry in the protein sequence database (dashed arrows).

as SEQUEST,[20] Mascot,[21] MS-Tag,[22] and Sonar[23] are used to assign peptides to MS/MS spectra. These programs compare each acquired MS/MS spectrum against those obtained from a sequence database and use various scoring schemes to find the best matching peptide. However, they are known to produce a significant number of incorrect peptide assignments.[24] The process of validating peptide assignments often relies on time-consuming manual verification. This data analysis bottleneck can be significantly reduced by adopting statistical models for validation of peptide assignments.[25]

The ultimate goal of inferring protein identities based upon peptide assignments remains a challenge, even when statistical models are employed for validating those assignments (Figure 1). One must initially group all assigned peptides according to their corresponding proteins in the database. This is particularly difficult when an assigned peptide is “degenerate”, in the sense that its sequence is present in more than a single entry in the protein sequence database. Such cases often result from the use of eukaryotic databases, which contain homologous and redundant entries, and make it difficult to infer the particular corresponding protein(s) present in the original sample.[26] Once grouping is complete, the assigned peptides corresponding to an individual protein, and their probabilities, must be combined to compute a single protein confidence measure that is effective at distinguishing the correct from incorrect protein identifications. A particular challenge in that regard is the detection of correct protein identifications with only a single corresponding assigned peptide in the data set, since the majority of incorrect protein identifications also have only one corresponding peptide.

Several software tools have been described that facilitate the identification of proteins based upon MS/MS data. Filtering and visualization programs such as INTERACT,[27] DTAselect,[28] and CHOMPER[29] simply report the list of proteins corresponding to the peptides assigned to MS/MS spectra, without attempting to resolve the cases of degenerate peptides or to estimate probability-based confidence measures. Mascot and Sonar group peptides according to their corresponding proteins and report a score for each protein intended to indicate the confidence of the identification. Qscore[30] estimates confidence levels of protein identifications from SEQUEST search results by taking into account the total number of identified peptides in the data set and the number of identified peptides corresponding to each protein. The effect of multiple peptides on the confidence of protein identifications was also described for a modified version of SEQUEST.[31] Though the scores provided by these tools can be used as criteria for filtering data in order to help separate correct from incorrect protein identifications, they provide no means to estimate the resulting false positive error rate (fraction of proteins passing the filter that are incorrect) and sensitivity (fraction of correct proteins passing the filter).

In this paper, we describe a model for computing accurate probabilities that proteins are present in a sample on the basis of peptides assigned to MS/MS spectra acquired from a proteolytic digest of the sample. This model has as its input a list of assigned peptides along with probabilities that those assignments are correct. Probabilities that peptide assignments are correct can be obtained, for example, according to the method described in ref 25, or any alternative method, as long as they are accurate. Furthermore, the model does not require peptide assignments to MS/MS spectra made by database search, but should be applicable to other computational approaches developed to analyze MS/MS spectra as well, such as those based on a combination of de novo sequencing and database search.[32−34] It computes a probability that a protein is present by combining together the probabilities that corresponding peptides are correct after adjusting them for observed protein grouping information. The model

[20] Eng, J. K.; McCormack, A. L.; Yates, J. R., III. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−989.

[21] Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551−3567.

[22] Clauser, K. R.; Baker, P.; Burlingame, A. L. *Anal. Chem.* **1999**, *71*, 2871−2882.

[23] Field, H. I.; Fenyo, D.; Beavis, R. C. *Proteomics* **2002**, *2*, 36−47.

[24] Keller, A. D.; Purvine, S.; Nesvizhskii, A. I.; Stolyar, S.; Goodlett, D. R.; Kolker, E. *OMICS* **2002**, *6* (2), 207−212.

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

[26] Rappsilber, J.; Mann, M. *Trends Biochem. Sci.* **2002**, *27*, 74−78.

[27] Han, D. K.; Eng, J.; Zhou, H.; Aebersold, R. *Nat. Biotechnol.* **2001**, *19*, 946−951.

[28] Tabb, D. L.; McDonald, W. H.; Yates, J. R. *J. Proteome Res.* **2002**, *1*, 21−26.

[29] Eddes, J. S.; Kapp, E. A.; Frecklington, D. F.; Connolly, L. M.; Layton, M. J.; Moritz, R. L.; Simpson, R. J. *Proteomics* **2002**, *2*, 1097−1103.

[30] Moore, R. E.; Young, M. K.; Lee, T. D. *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 378−386.

[31] MacCoss, M. J.; Wu, C. C.; Yates, J. R. *Anal. Chem.* **2002**, *74*, 5593−5599.

[32] Taylor, J. A.; Johnson, R. S. *Rapid Commun. Mass Spectrom.* **1997**, *11*, 1067−1075.

[33] Mann, M.; Wilm, M. *Anal. Chem.* **1994**, *66*, 4390−4399.

[34] Shevchenko, A.; Sunyaev, S.; Loboda, A.; Shevchenko, A.; Bork, P.; Ens, W.; Standing, K. G. *Anal. Chem.* **2001**, *73*, 1917−1926.

Table 1. Experimental Data Sets Used in the Study

data set	sample	database used for search	no. of LC/MS/MS runs	no. of MS/MS spectra[f]		
				1+	2+	3+
18prot_Hinf	18-protein mix[a]	*H. influenzae*[d] + 18 proteins[e]	22	504	18 496	18 044
18prot_Dr	18-protein mix[a]	*Drosophila*[d] + 18 proteins[e]	22	422	18 368	18 022
18prot_Hum	18-protein mix[a]	human[d] + 18 proteins[e]	22	401	18 180	17 824
18prot.sub1_Hum	18-protein mix[a]	*H. influenzae*[d] + 18 proteins[e]	1	26	694	664
18prot.sub4_Hum	18-protein mix[a]	*H. influenzae*[d] + 18 proteins[e]	4	13	1 294	12 95
Hinf_Hum	*H. influenzae*[b]	human[d] + *H. influenzae*[d]	15	2006	15 791	15 553
Halo_Hum	*Halobactereum*[c]	human[d] + *Halobacterium*[d]	5	1834	5 829	5 285

[a] Sample composed of 18 highly purified proteins (from bovine, chicken, rabbit, *E. coli*, *S. cerevisiae,* and *B. lichenformis*).[24] [b] *H. influenzae* membrane fraction sample.[43] [c] *Halobacterium*, soluble fraction sample.[44] [d] Protein sequence databases extracted from ref 35. [e] Sequences of the 18 purified proteins and common sample contaminants such as keratin.[24] [f] Number of MS/MS database searches performed on $[M + H]^+$, $[M + 2H]^{2+}$, and $[M + 3H]^{3+}$ spectra.

handles cases when assigned peptides are degenerate by apportioning each such peptide among all its corresponding proteins in order to derive a minimal protein list sufficient to account for the observed peptide assignments. Furthermore, the model collapses redundant database entries into a single identification and groups together those proteins that are impossible to differentiate on the basis of peptides assigned to MS/MS spectra. We evaluate the model using data sets of MS/MS spectra generated from a sample of 18 purified proteins,[24] as well as complex *Haemophilus influenzae* and *Halobacterium* samples. We demonstrate that it produces accurate protein probabilities with high power to discriminate between correct and incorrect protein identifications, including those corresponding to only a single assigned peptide in the data set.

This method for estimating the probability that a particular protein is present in the sample given the acquired mass spectrometric information is of great importance to proteomics. It is fully automated and fast and does not rely on subjective "expert" judgment (manual validation). It allows filtering of large-scale data sets with predictable sensitivity and false positive identification error rates and provides a standardized way for publishing large-scale proteomics data sets in the literature. Finally, it makes the data analysis consistent and transparent, providing a way to compare results of different experimental groups obtained using different experimental protocols, different mass spectrometers, and even different MS/MS database search tools.

EXPERIMENTAL DATA SETS

A description of all experimental data sets used in this study is given in Table 1. All MS/MS spectra were generated using a similar experimental protocol. Protein samples were proteolyzed with trypsin and analyzed by LC/MS on an ESI-ITMS (ThermoFinnigan, San Jose, CA) using a top-down data-dependent ion selection approach.[36] The spectra were searched with the SEQUEST program.[20] All resulting peptide assignments in the data set were analyzed using PeptideProphet, a software tool implementing the statistical model described in ref 25, improved to include the number of missed cleavages[37] in the sequence of the assigned peptide, and extended to analyze peptide assignments to spectra of $[M + H]^+$ ions (see Supporting Information). The peptide probabilities computed by PeptideProphet have been shown to be very discriminative and accurate and were used as input to the protein statistical model described in this work. It should be noted that the protein statistical model can accept as input a list of probabilities corresponding to each peptide assignment regardless of how these probabilities were computed, as long as they are accurate.

RESULTS AND DISCUSSION

Peptide Probability Estimates. Since MS/MS spectra are produced by peptides, and not by proteins, all conclusions about what proteins are present in the sample are based upon the identification of peptides that correspond to them. Thus, estimation of the probability that a particular protein is present in the original sample can be facilitated by having a statistical model for validation of the identifications made at the peptide level. In ref 25, a robust statistical approach was presented for estimation of the accuracy of peptide assignments to MS/MS spectra made by database search algorithms. This approach is based on the use of the expectation−maximization (EM) algorithm to derive a mixture model of correct and incorrect peptide identifications from the data. Observed data (denoted as D) includes available information regarding database search results, such as database search scores and properties of the assigned peptides, that help to distinguish the correct (denoted as "+") and incorrect (denoted as "−") peptide assignments in the data set. By employing the observed information about each peptide assignment, the method learns to distinguish correctly from incorrectly assigned peptides in the data set and computes for each peptide assignment to a spectrum a probability of being correct, $p(+|D)$, using Bayes' Law:

$$p(+|D) = \frac{p(D|+)p(+)}{p(D|+)p(+) + p(D|-)p(-)} \tag{1}$$

where $p(D|+)$ and $p(D|-)$ are the probabilities of an assigned peptide to an MS/MS spectrum having information D among correctly and incorrectly assigned peptides, respectively, and $p(+)$ and $p(-)$ are prior probabilities of a correct and incorrect peptide assignment, respectively. The prior probabilities are the overall proportions of correct and incorrect peptide assignments in the

[35] Extracted from ftp://ftp.ncicrf.gov/pub/nonredun/protein.nrdb.Z.

[36] Goodlett, D. R.; Keller, A.; Watts, J. D.; Newitt, R.; Yi, E. C.; Purvine, S.; Eng, J. K.; von Haller, P.; Aebersold, R.; Kolker, E. *Rapid Commun. Mass Spectrom.* **2001**, *15*, 1214−1221.

[37] Parker, K. C. *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 22−39.

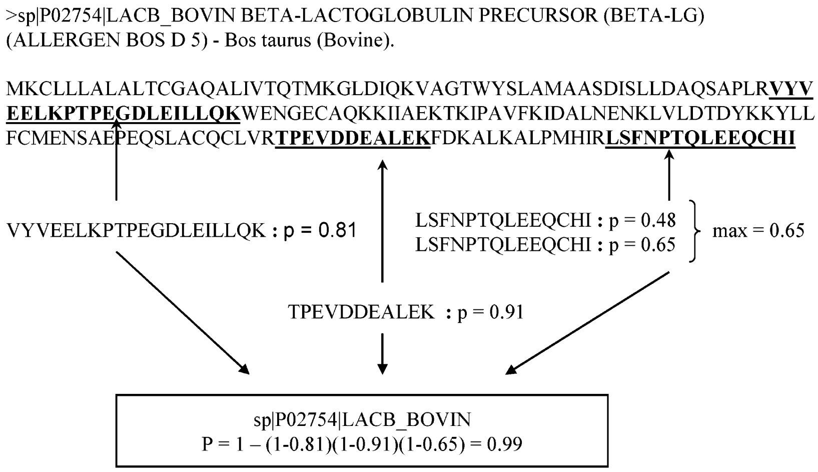

>sp|P02754|LACB_BOVIN BETA-LACTOGLOBULIN PRECURSOR (BETA-LG)
(ALLERGEN BOS D 5) - Bos taurus (Bovine).
MKCLLLALALTCGAQALIVTQTMKGLDIQKVAGTWYSLAMAASDISLLDAQSAPLRVYV
EELKPTPEGDLEILLQKWENGECAQKKIIAEKTKIPAVFKIDALNENKLVLDTDYKKYLL
FCMENSAEPEQSLACQCLVRTPEVDDEALEKFDKALKALPMHIRLSFNPTQLEEQCHI
VYVEELKPTPEGDLEILLQK : p = 0.81
LSFNPTQLEEQCHI : p = 0.48
LSFNPTQLEEQCHI : p = 0.65
max = 0.65
TPEVDDEALEK : p = 0.91
sp|P02754|LACB_BOVIN
P = 1 – (1-0.81)(1-0.91)(1-0.65) = 0.99

Figure 2. Illustration of how eq 3 is used to compute protein probabilities. A protein is identified by several distinct peptides assigned to MS/MS spectra. Probabilities that the peptide assignments are correct are combined together to estimate the probability that the corresponding protein is present in the sample. Assignments of the same peptide to multiple MS/MS spectra make a single contribution with the maximum probability of all instances of that peptide in the data set.

data set and can be considered as a measure of data quality. The parameters governing the distribution of database search scores and other information, as well as the prior probabilities, are learned from the data itself. This ensures that this method is robust toward variations in sample purity, mass spectral quality, proteolytic digest efficiency, and other factors.[25]

Protein Probability Estimates. Accurate probabilities that peptide assignments are correct, computed for example by the mixture model EM method described above, can be used to estimate the probability that a particular protein is present in the original sample. In general, there can be many distinct peptides assigned to spectra, each of which corresponds to the same particular protein of interest. Furthermore, each distinct peptide may be assigned to more than a single spectrum in the data set. Probabilities that peptide assignments are correct vary for different peptides and even for repeated observation of the same peptide. These assigned peptides each contribute evidence for the presence of the corresponding protein. For the sake of simplicity, this section describes an analysis that neglects all peptide assignments that correspond to more than a single entry in the protein sequence database, since in those cases their true corresponding proteins are ambiguous.

If each peptide assignment to a spectrum is considered independent evidence for its corresponding protein, then the probability P that a protein is present in the sample can be computed as the probability that at least one peptide assignment corresponding to the protein is correct,

$$P = 1 - \prod_i \prod_j (1 - p(+|D_i^j)) \qquad (2)$$

where each distinct peptide i corresponds to the protein of interest and $p(+|D_i^j)$ is the computed probability that the jth assignment of peptide i (its peptide assignment information denoted D_i^j) to a spectrum in the data set is correct. This formula likely overestimates the probability, however, since assignments of the same peptide to multiple spectra are not justifiably independent events when those spectra have nearly identical fragmentation patterns, as is often the case.[21,30] For example, multiple spectra corresponding to a peptide that is not in the database, perhaps due to a posttranslational modification, would each likely be assigned the same incorrect peptide, leading to a misleadingly high probability for the corresponding protein. To illustrate this, eq 2 was applied to the data set of 22 LC/MS/MS runs generated from a mixture of 18 known proteins and searched against the *Drosophila* sequence database, 18prot_Dr (see Table 1 and also ref 25). As a result, four incorrect protein identifications were assigned a probability 0.99 or greater, and eight incorrect protein identifications were assigned a probability 0.9 or greater, all based on a single corresponding peptide observed several times in the data set. For example, the same peptide corresponding to incorrect protein GP:AY010604_1 was assigned to three different spectra in the data set, with probabilities of being correct 0.42, 0.52, and 0.64, respectively. Using eq 2, GP:AY010604_1 was then assigned a combined protein probability of 0.9. This inappropriately high computed probability suggests that multiple identifications of the same peptide in a data set should not result in increased confidence that the corresponding protein is correct.

A more conservative estimate of the probability that a protein is present in the sample can be computed as

$$P = 1 - \prod_i (1 - \max_j p(+|D_i^j)) \qquad (3)$$

using only a single contribution for all j assignments to spectra of each distinct peptide i that corresponds to that protein, the contribution determined by the maximum probability for all assignments of that peptide, $\max_j p(+|D_i^j)$. An illustration of how eq 3 is used to compute protein probabilities is shown in Figure 2. This conservative approach results in more accurate protein probabilities and is implemented in the model. For example, using eq 3 in place of eq 2, the computed probability of the incorrect protein identification GP:AY010604_1 described above was reduced from 0.9 to 0.64, and overall, no probabilities of incorrect proteins were 0.99 or greater, and three fewer were 0.9 or greater.

For the purpose of clarity, in all subsequent discussions it will be assumed that all assignments to spectra of each distinct peptide are represented by a single contribution having maximum probability. Note, however, that unlike the situation of multiple assignments of a peptide to MS/MS spectra of the same precursor ion charge state, assignments corresponding to the same peptide but with different charge state all contribute to the evidence for the presence of the corresponding protein. This conclusion is supported by the observation that peptides with different charge state have significantly different MS/MS fragmentation patterns,[38] and it is rare to observe an incorrect peptide assigned with significant probabilities to spectra of multiple precursor ion charge states.

Adjusting Peptide Probabilities for Observed Protein Grouping Information. The grouping of peptides according to their corresponding proteins in eq 3 provides valuable new information regarding the validity of the peptide assignments. That is because correct peptide assignments, more than incorrect ones, tend to correspond to "multihit" proteins, those to which other correctly assigned peptides correspond.[39] In contrast, incorrect peptide assignments tend to correspond to proteins to which no other correctly assigned peptide corresponds. This trend is particularly pronounced for "high coverage" data sets, i.e., data sets consisting of a relatively large number of acquired MS/MS spectra with respect to the complexity of the sample (number of proteins in the sample). As a result, even when computed probabilities of correct peptide assignments are accurate in the context of the complete data set for which they are calculated, they may not be as accurate for subsets of peptides grouped according to corresponding protein. For example, in the case of "multihit" proteins, greater than the expected half of all corresponding assigned peptides with computed probability 0.5 are likely to be correct. Consequently, before applying eq 3, computed peptide probabilities must be made accurate for subsets of peptides grouped according to corresponding proteins by adjusting them to reflect whether the protein is multihit in the data set.

A measure of whether a peptide corresponds to a multihit protein in a data set is its estimated number of sibling peptides (NSP), defined as the expected number of other correctly identified peptides that correspond to the same protein. The NSP value for peptide i, NSP$_i$, is computed as the sum of probabilities of correct assignments to spectra of the other peptides that correspond to the same protein:

$$\mathrm{NSP}_i = \sum_{\{m|m\neq i\}} p(+|D_m) \qquad (4)$$

where peptide m is another distinct peptide corresponding to the protein of interest, and $p(+|D_m)$ is the maximum probability of all assignments of peptide m in the data set. For example, in the situation shown in Figure 2, peptide VYVEELKPTPEGDLEILLQK has an NSP value 1.56 (calculated as 0.91 + 0.65, the sum of the probabilities of its sibling peptides, TPEVDDEALEK and LSFNPTQLEEQCHI).

The difference in NSP values between correct and incorrect peptide assignments is particularly pronounced for the data sets of 22 LC/MS/MS runs generated from a sample containing only 18 control proteins, which, given the number of acquired spectra, should be considered a high coverage data set. For example, in the 18prot_Dr data set, the majority (92%) of correct peptide assignments have NSP values above 5, with average value around 7, reflecting the fact that all of the identified control proteins were identified by multiple peptides. In contrast, fewer than 1% of the incorrect peptide assignments, namely, those that are chance hits to one of the control proteins,[24] have NSP values above 5, and the majority of incorrect assignments have NSP values below 0.25, with the average value of 0.01. The difference between NSP distributions among correct and incorrect peptide assignments is expected to be somewhat smaller, though still significant, for lower coverage data sets such as those typically produced from complex protein samples.

The probabilities that database search results are correct based upon peptide assignment information, D, $p(+|D)$, can be adjusted to take into account NSP making the reasonable assumption that NSP is independent of the parameters included in D, i.e., database search scores, number of tryptic termini, and number of missed cleavages, among correctly and incorrectly assigned spectra:

$$p(+|D,\mathrm{NSP}) = \frac{p(+|D)p(\mathrm{NSP}|+)}{p(+|D)p(\mathrm{NSP}|+) + p(-|D)p(\mathrm{NSP}|-)} \qquad (5)$$

where $p(+|D,\mathrm{NSP})$ is the probability that peptide assignment is correct given its D and its estimated number of sibling peptides, NSP, and $p(\mathrm{NSP}|+)$ and $p(\mathrm{NSP}|-)$ are the probabilities of having a particular NSP value according to the distribution of correct or incorrect peptide assignments, respectively. For simplicity, NSP values are made discrete by binning. The probability that a correctly assigned peptide has an NSP value in bin k can then be computed by summation over only those peptides with NSP value in bin k:

$$p(\mathrm{NSP}|+) = \frac{1}{Np(+)} \sum_{\{i|\mathrm{NSP}_i \in k\}} p(+|D_i,\mathrm{NSP}_i) \qquad (6)$$

where N is the total number of peptide assignments and $p(+)$ is the prior probability of a correct peptide assignment, computed by summation over all peptides i:

$$p(+) = \frac{1}{N}\sum_i p(+|D_i,\ \mathrm{NSP}_i) \qquad (7)$$

The NSP distribution among incorrect peptide assignments is computed in an analogous manner. The adjusted probabilities that peptide assignments to MS/MS spectra are correct, given by eq 5, have improved power to discriminate correct and incorrect database search results (See Supporting Information.). Furthermore, they are more accurate among subsets of peptides grouped according to corresponding protein and thus suitable for substitution into eq 3 to compute probabilities that those proteins are present in the sample.

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

[39] Choudhary, J. S.; Blackstock, W. P.; Creasy, D. M.; Cottrell, J. S *Proteomics* **2001**, *1*, 651−667.

NSP distributions are expected to vary from data set to data set, reflecting a number of parameters such as data set size, protein sequence database size, number of proteins in the original sample and their relative concentrations, and data quality. The NSP distributions can be derived from each data set as a mixture model using the EM algorithm.[40] Initially, the unadjusted peptide probabilities $p(+|D)$ are used to compute the estimated number of sibling peptides, NSP$_i$, for each peptide assignment i in the data set (eq 4). Applying the EM algorithm, adjusted probabilities are then computed according to eq 5 alternatively with mixture model NSP distributions among correctly and incorrectly assigned spectra, $p(\text{NSP}|+)$ and $p(\text{NSP}|-)$, respectively, according to eqs 6 and 7, until a fixed point is reached. The resulting adjusted peptide probabilities $p(+|D,\text{NSP})$ are then substituted into eq 3 to compute protein probabilities. This multistep approach facilitates the robust analysis of alternative groupings of data. Initial probabilities that peptide assignments are correct based upon peptide assignment information D can be computed independently for disparate data sets (e.g., spectra collected from different samples or using different mass spectrometers) and then adjusted to take into account NSP only in the final combined data set for which protein probabilities are desired.

Figure 3 illustrates how NSP distributions vary depending on sample complexity and data set size. Figure 3A plots the logarithm of the ratio $p(\text{NSP}|+)/p(\text{NSP}|-)$ learned by the model in each NSP bin k for the 18prot_Hum and Hinf_Hum data sets. A ratio greater than unity (positive logarithm) indicates that the probabilities are boosted by including NSP information, whereas a ratio less than unity (negative logarithm) indicates that the NSP adjustment reduces the probability that a peptide assignment is correct. These two data sets have approximately the same number of spectra searched against databases of nearly identical size, yet whereas the 18prot_Hum data set was generated from a sample with only 18 proteins, the Hinf_Hum data set was generated from a complex sample containing hundreds of proteins. Thus, one would expect a greater percent of correctly identified proteins in the former to be multihit. Indeed, log $p(\text{NSP}|+)/p(\text{NSP}|-)$ for the 18prot_Hum data set is more strongly positive at high NSP bins relative to the Hinf_Hum data set and more strongly negative at low NSP bins (Figure 3A). As a result, incorporating NSP information boosts probabilities of peptides with high NSP values (corresponding to multihit proteins) and penalizes those with low NSP values to a greater degree in the case of the 18prot_Hum data set than in the case of the Hinf_Hum data set. NSP distributions are also affected by the number of spectra in the data set. As the number of MS/MS spectra generated from a protein sample increases, more of the correctly identified proteins will be multihit. Figure 3B plots log $p(\text{NSP}|+)/p(\text{NSP}|-)$ learned by the model for three data sets of increasing size, 18prot.sub1_Hum (1 LC/MS/MS run), 18prot.sub4_Hum (4 runs), and 18prot_Hum (all 22 runs), all generated from the same sample of 18 purified proteins. It is expected that as the size of the data set increases with the sample complexity kept constant, sample coverage increases whereby more correctly identified proteins are multihit. As a result, the amount of peptide probability adjustment for peptides with low or very high NSP increases with increasing data set size (Figure 3B). The effect of data set size on the learned NSP distributions is expected to be more significant in the case of protein samples of higher complexity than the 18-protein mix.

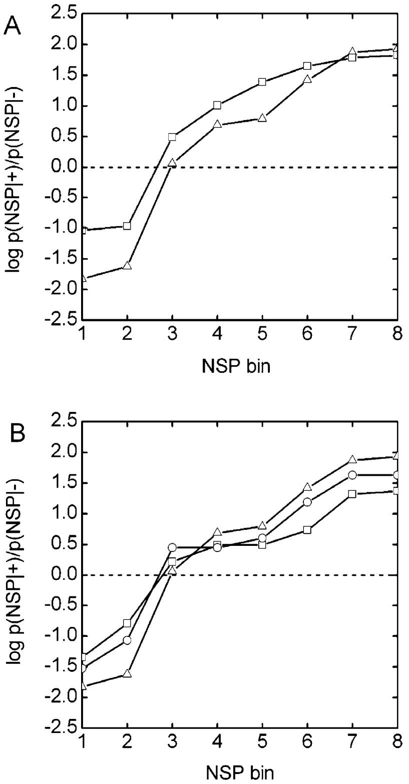


Figure 3. Dependence of NSP (expected number of sibling peptides) distributions on sample complexity and data set size. (A) The logarithm of the ratio $p(\text{NSP}|+)/p(\text{NSP}|-)$ learned by the model for each NSP bin k for comparable numbers of spectra generated either from a low-complexity sample, 18prot_Hum, (triangles) or from a high-complexity sample, Hinf_Hum, (squares). (B) Same for three data sets of increasing size, 18prot.sub1_Hum, 1 LC/MS/MS run (squares), 18prot.sub4_Hum, 4 runs (circles), and 18prot_Hum, 22 runs (triangles), each generated from the same sample. NSP bins k are defined as $0 \le \text{NSP} < 0.1$ (bin 0), $0.1 \le \text{NSP} < 0.25$ (bin 1), $0.25 \le \text{NSP} < 0.5$ (bin 2), $0.5 \le \text{NSP} < 1$ (bin 3), $1 \le \text{NSP} < 2$ (bin 4), $2 \le \text{NSP} < 5$ (bin 5), $5 \le \text{NSP} < 15$ (bin 6), and $15 \le \text{NSP}$ (bin 7).

For each peptide, the amount of adjustment for NSP should, in principle, also depend on the likelihood of observing a particular number of its sibling peptides. For any protein, the likelihood of observing a particular number of peptides depends on a number of factors such as its abundance in the sample, length, the number of expected tryptic peptides (if proteins are digested using trypsin) or, in case of ICAT experiments,[41] the number of cysteine-containing peptides. In addition, some peptides are rarely, if ever, identified using mass spectrometry methods because their physicochemical properties result in poor ionization efficiency or incomplete fragmentation. As a result, some proteins might be expected to produce only one distinct peptide that could possibly be identified in an experiment. Such peptides, if indeed observed, should therefore not have their probability of being correct adjusted downward due to a low NSP value (no sibling peptides), even if the majority of other proteins are identified by multiple distinct peptides. Note, however, that the amount of adjustment

[40] Dempster, A.; Laird, N.; Rubin, M. *J. R. Stat. Soc.* **1977**, *B39* (1), 1−38.

[41] Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R. *Nat. Biotechnol.* **1999**, *17*, 994−999.

for having a low NSP value depends not only on the ratio $p(\mathrm{NSP}|+)/p(\mathrm{NSP}|-)$ learned by the model but also on the unadjusted peptide probability $p(+|D)$; see eq 5. As a result, among all peptides in the data set having a low NSP value, peptides with the unadjusted probability $p(+|D)$ close to 1 are penalized to a lower degree then those in the intermediate- or low-probability range. This ensures that, in practice, adjustment for NSP does not result in the loss of protein identifications based on only one or two distinct peptides corresponding to them, which is often the case with very small or low-abundant proteins, as long as they are identified by at least one *high-probability* peptide.

Degenerate Peptides. Peptides assigned to MS/MS spectra that correspond to more than a single entry in the protein sequence database are "degenerate" and present a challenge since their true corresponding proteins are uncertain. Such cases most often result from the presence of homologous proteins, splicing variants, or redundant entries in the protein sequence database.[26] Degenerate peptides are more prevalent with searches using large databases. For example, they routinely comprise 20% or more of all peptides assigned to MS/MS spectra upon searching a human sequence database.

The most likely protein(s) corresponding to degenerate peptides can be inferred by apportioning such peptides among their possible corresponding proteins according to the estimated probabilities of those proteins in the sample, while in turn computing the protein probabilities taking into account those estimated apportionments. For the sake of clarity, this section describes an analysis neglecting NSP information. The full analysis combining treatment of degenerate peptides and adjustment of peptide probabilities for NSP is presented in the next section. If peptide i corresponds to N_s different proteins, then the relative weight, w_i^n, that this peptide actually corresponds to protein n ($n = 1 \ldots N_s$) is determined according to the probability of protein n relative to those of all N_s proteins:

$$w_i^n = \frac{P_n}{\sum_{s=1\ldots N_s} P_s} \tag{8}$$

In turn, the protein probabilities are computed according to eq 3 taking into account the weights as well as the peptide probabilities:

$$P_n = 1 - \prod_i (1 - w_i^n p(+|D_i)) \tag{9}$$

where the contribution of peptide i is weighted by its estimated apportionment to protein n, w_i^n, and $p(+|D_i)$ is the maximum probability among multiple assignments of peptide i to spectra in the data set, $\max_j p(+|D_i^j)$. The model learns the degenerate peptide weights iteratively using an EM-like algorithm. Peptides are initially equally apportioned among their possible corresponding proteins, and the protein probabilities calculated according to eq 9. Then, both eqs 8 and 9 are applied iteratively until a fixed point is reached.

The weights learned by the model reflect the likelihood that each of the proteins corresponding to degenerate peptides is present in the sample. Equation 8, requiring that the weights for each degenerate peptide sum to unity, is based upon the supposition that each degenerate peptide has only one corresponding protein in the original sample. This is aimed at deriving the simplest list of proteins sufficient to explain the observed peptides assigned to MS/MS spectra in the data set and can be described in essence as Occam's razor, "plurality should not be posited without necessity".[42] For example, Figure 4 shows a common situation in which a degenerate peptide, with a probability $p(+|D) = p_1$ of being a correct identification, corresponds to two different proteins, A and B. There are no peptide identifications in the data set corresponding to protein A and not protein B. On the other hand, protein B has nondegenerate evidence in the form of peptide 2, with a probability $p(+|D) = p_2$, which corresponds only to protein B. Initially, both proteins are assumed equally likely, $P_A = P_B$, and the weights w_1^A and w_1^B are equally apportioned according to eq 8. The weight w_2^B is equal to 1 and fixed at that value throughout all iterations since peptide 2 contributes to no proteins other than B. Substitution of the weights and peptide probabilities in eq 9 updates protein probabilities and, due to presence of peptide 2, P_B becomes greater than P_A, and consequently, w_1^B becomes greater than w_1^A. The iterations continue until a fixed point is reached. In this particular case, the solution given by eqs 8 and 9 is that protein B acquires all the weight from peptide 1, i.e., $w_1^B = 1$ and $w_1^A = 0$, so that $P_A = 0$ and $P_B = 1 - (1 - p_1)(1 - p_2)$. In the above example, any nondegenerate evidence for protein B, regardless of how low its probability, would result in the same apportionment of $w_1^B = 1$ and $w_1^A = 0$. Since in any realistic data set there is a large number of incorrect peptide identifications with low probabilities, i.e., chance assignments of peptides corresponding to randomly selected proteins in the database, improved performance of the model was achieved by introducing an empirically selected minimum probability threshold of 0.2 for peptides used in eq 9 to compute protein probabilities. It should be noted that, in large data sets, the network of degenerate peptides and their corresponding proteins becomes quite complex relative to the simple example illustrated in Figure 4. Such complexity, however, presents no problem for the approach described above.

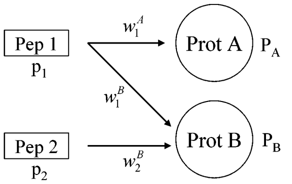


Figure 4. Illustration of a degenerate peptide case. Peptide 1, having probability p_1 of being a correct identification, corresponds to two different proteins, A and B. Protein B has nondegenerate evidence in the form of peptide 2 with probability p_2, which corresponds to protein B, and not A. Apportionments of peptide 1 among the two proteins, w_1^A and w_1^B, and the protein probabilities, P_A and P_B, are learned iteratively using an EM-like algorithm.

It is not uncommon, even in organisms with small genomes, to encounter a situation when several database entries share a set of observed peptides and are in essence indistinguishable given the available mass spectrometric data. Such entries may include various isoforms or splicing variants that could be products of the same gene or products of different but related genes. In addition, they may simply reflect a significant number of redun-

[42] Good, I. J. *Proc. R. Soc. London A* **1977**, *354*, 303−330.

dancies present in many databases due to incomplete protein sequences or sequencing errors. To facilitate interpretation of the model results, the indistinguishable proteins are reported together in a group and assigned a single probability that any member is present in the sample.

Combined Treatment of Degenerate Peptides and NSP. Inclusion of NSP and treatment of degenerate peptide cases can be combined in a single step. If peptide i is shared between N_s different proteins, then that peptide would have N_s different NSP values, one for each protein n ($n = 1 \ldots N_s$) computed by summing over other peptides m also corresponding to that protein

$$\mathrm{NSP}_i^n = \sum_{\{m|m \neq i\}} w_m^n p(+|D_m) \qquad (10)$$

The NSP distributions among correct and incorrect peptide assignments, $p(\mathrm{NSP}|+)$ and $p(\mathrm{NSP}|-)$, are calculated for each bin k in a similar way:

$$p(\mathrm{NSP}|+) = \frac{1}{Np(+)} \sum_n \sum_{\{i|\mathrm{NSP}_i^n \in k\}} w_i^n p(+|D_i,\mathrm{NSP}_i^n) \qquad (11)$$

where N is the total number of peptide assignments and $p(+)$ is the prior probability of a correct peptide assignment. The protein probabilities are computed according to eq 9 using peptide probabilities adjusted for NSP:

$$P_n = 1 - \prod_i (1 - w_i^n p(+|D_i,\mathrm{NSP}_i^n)) \qquad (12)$$

Initially, estimates of the weights w_i^n are calculated using eq 8, after which NSP values are computed using eq 10. The NSP distributions are then estimated according to eq 11, and the peptide and protein probabilities according to eq 5 and eq 12 iteratively until a fixed point is reached. Note that if a peptide is shared between N_s different proteins and, as a result, has N_s different NSP values, then that peptide gives as many contributions to the negative and positive NSP distributions. Upon termination of the algorithm, the model learns accurate NSP distributions reflecting the prevalence of multihit proteins in the data set and the proteins corresponding to degenerate peptides that are more likely to be present in the sample, given the Occam's razor constraint. Finally, it computes an accurate probability that each protein is present in the sample.

Evaluation of the Model. To evaluate the performance of the statistical model, it was first applied to data sets generated from a sample with 18 known proteins. Table 2 shows the number of correct and incorrect protein identifications with probability greater than or equal to 0.7 computed with or without adjusting peptide probabilities to account for NSP. The first three data sets, 18prot_Hinf, 18prot_Dr, and 18prot_Hum, were generated from the same data set of MS/MS spectra, yet searched against databases of increasing size (Table 1). The 18prot_Hum data set has a significant number of degenerate peptides, since the human sequence database contains many homologous proteins, splicing variants, and redundant entries. By comparison, the 18prot_Dr data set has very few, and the 18prot_Hinf data set, even fewer, degenerate peptides. Table 2 shows that the model performed well on all data sets, regardless of the database used in the search. Employing NSP information, it produced 18 correct protein identifications and only 1 (18prot_Hum) or 0 (18prot_Hinf, 18prot_Dr) incorrect identifications with computed probability 0.7 or greater. In contrast, when NSP was *not* employed to adjust peptide probabilities, the number of incorrect proteins assigned probabilities 0.7 or greater increased significantly (34 in the case of 18prot_Hum).

Table 2. Evaluation of Model Performance on Data Sets Generated from the Mixture of 18 Purified Proteins[a]

		no. of protein identifications			
		$P \geq 0.7$		$P \geq 0.9$	
data set	model	+	−	+	−
18prot_Hinf	with NSP	18	0	18	0
	without NSP	19	26	19	8
18prot_Dr	with NSP	18	0	18	0
	without NSP	19	29	19	12
18prot_Hum	with NSP	18	1	18	1
	without NSP	18	34	18	12
18prot.sub1_Hum	with NSP	9	0	8	0
	without NSP	9	2	9	1
18prot.sub4_Hum	with NSP	12	0	12	0
	without NSP	12	5	12	2

[a] The numbers of correct (+) and incorrect (−) protein identifications having computed probability (P) equal to or greater than the indicated minimum probability thresholds are shown. Since the samples also contained several common contaminants such as keratin, the total number of protein identifications considered to be correct can exceed 18.

Table 2 also shows the results obtained for different data set sizes. The model was applied to subsets of the 18prot_Hum data set containing data from only 1 (18prot.sub1_Hum) or 4 (18prot.sub4_Hum) LC/MS/MS runs, as well as to the entire data set of 22 runs (Table 1). The model learned different NSP distributions in these three data sets (Figure 3) and in each case employed the learned distributions to compute accurate and discriminating protein probabilities. In all three cases, many more correct than incorrect proteins were assigned probabilities 0.7 or greater. Similar results were obtained for the number of correct and incorrect proteins with computed probability greater than or equal to 0.9.

To evaluate the accuracy and discriminating power of computed protein probabilities for more realistic data sets generated from samples with large numbers of proteins, the analysis was applied to database search results of MS/MS spectra generated from complex *H. influenzae*,[43] Hinf_Human, and *Halobacterium*,[44] Halo_Hum, samples. All MS/MS spectra were searched against a human protein sequence database appended with the much smaller database of the corresponding sample organism (see Table 1 for details). Figure 5 shows the accuracy of the protein probabilities computed by the model. All protein identifications were sorted according to the computed probability that those identifications are correct, and the actual probabilities (fraction

(43) Kolker, E.; Purvine, S.; Galperin, M. Y.; Stolyar, S.; et al., submitted to *J. Bacter.*

(44) Ng, V. et al. Manuscript in preparation.

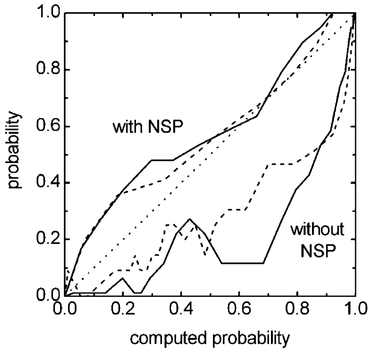


Figure 5. Accuracy of computed protein probabilities. The actual probability (fraction of protein identifications that are correct) among identifications with indicated computed probabilities derived from the *H. influenzae* sample, Hinf_Hum (solid line), and the *Halobacterium* sample, Halo_Hum (dashed line). Also shown is the accuracy of probabilities computed without adjusting peptide probabilities to account for protein grouping information (NSP). The expected probability for an ideal model is indicated by the dotted 45° line.

of correct proteins) were determined within a sliding window of 20 identifications. All identifications of human proteins, with the exception of common contaminants such as keratin, were considered to be incorrect, whereas all identifications of the proteins of the sample organism were inferred to be correct, with the exception of those due to chance peptide assignments.[24] The frequency of chance assignments was determined empirically as a ratio of the number of assignments to proteins of the sample organism with computed probability close to 0, to the total number of incorrect assignments in the same probability range, and was found to be around 4% for the Hinf_Hum data set and 7% for the Halo_Hum data set. Figure 5 shows that the protein probabilities calculated by the model are accurate (results of an ideal model are represented by the 45° line). Figure 5 also plots the accuracy of protein probabilities calculated without adjustment of peptide probabilities for NSP, which are significantly overestimated, especially in the range of intermediate probabilities. For example, in the Hinf_Hum data set, there were 36 proteins having a probability computed without adjustment for NSP between 0.4 and 0.6, when only 6 of those proteins (17%) were actually present in the sample. At the same time, the peptide probabilities used as an input in the model were shown to be accurate, if slightly underestimated (see Supporting Information). The overestimation of protein probabilities calculated without adjusting peptide probabilities to account for NSP can be explained by noting that nearly all of the 36 proteins not present in the sample were identified (incorrectly) on the basis of only one peptide having a significant probability of being correct. NSP adjustment penalizes just such peptides with low NSP values, resulting in more accurate protein probabilities, with 5 out the 12 proteins (42%) having assigned probabilities (computed with adjustment for NSP) between 0.4 and 0.6 being present in the sample.

The discriminating power of computed protein probabilities is illustrated in Figure 6A, which plots for the Hinf_Hum data set the false positive error rate versus sensitivity resulting from filtering data on the basis of various minimum computed probability thresholds. Each point along the curve represents the results of using a different filter (minimum probability threshold) to accept all protein identifications with computed probabilities at least as great. The results indicate that the probabilities computed by the model have high power to discriminate the correct protein identifications from the incorrect ones. For example, employing a minimum probability threshold of 0.7 yields 94% sensitivity (240 correct protein identifications) with a false positive error rate of 1.2% (3 incorrect identifications), which is close to 100% sensitivity (255 correct protein identifications) and 0% error rate (0 incorrect identifications) expected in this data set from an ideal filter. Interestingly, in this data set, 39% of all correct identifications passing the 0.7 filter (95 correct identifications) had only one peptide corresponding to them. Traditionally, these proteins, which often include low-abundance and low molecular weight proteins, are among the most challenging identifications and would be lost using suggested filtering criteria requiring two or more corresponding peptides.[30] Figure 6A also demonstrates that filtering the data on the basis of protein probabilities computed without adjustment for NSP results in lower sensitivity (and thus, number of correct protein identifications) for any given false positive error rate. Similar results were observed for the Halo_Hum data set. Accurate computed probabilities can be used to compute the estimated number of correct protein identifications in the data set (by summing the probabilities of all protein identifications in the data set). They can also be used predict the false positive error rate and sensitivity resulting from

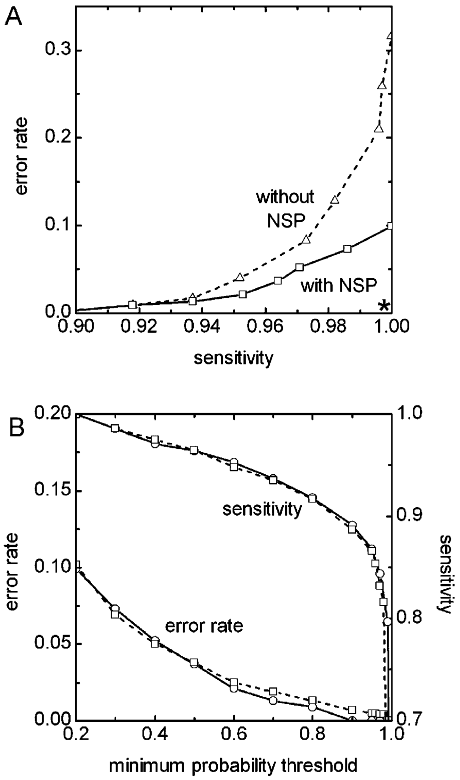


Figure 6. Sensitivity and false positive identification error rates using minimum computed protein probability thresholds. (A) Sensitivity/error rate tradeoff employing thresholds based upon probabilities computed with (solid line) or without (dashed line) adjustment for protein grouping information (NSP) for the *H. influenzae* sample, Hinf_Hum. The result of using an ideal filter (100% sensitivity and 0% error rate) is indicated by an asterisk. (B) Observed (solid line) and model predicted (dashed line) sensitivity and error rate as a function of minimum computed probability threshold derived for the same data set.

the use of any minimum probability threshold data filter.[25] Figure 6B shows for the Hinf_Hum data set that there is good agreement between the actual sensitivity and error rates and those predicted by the model. Thus, researchers can choose minimum probability thresholds that confer a desired sensitivity or error rate for any data set.

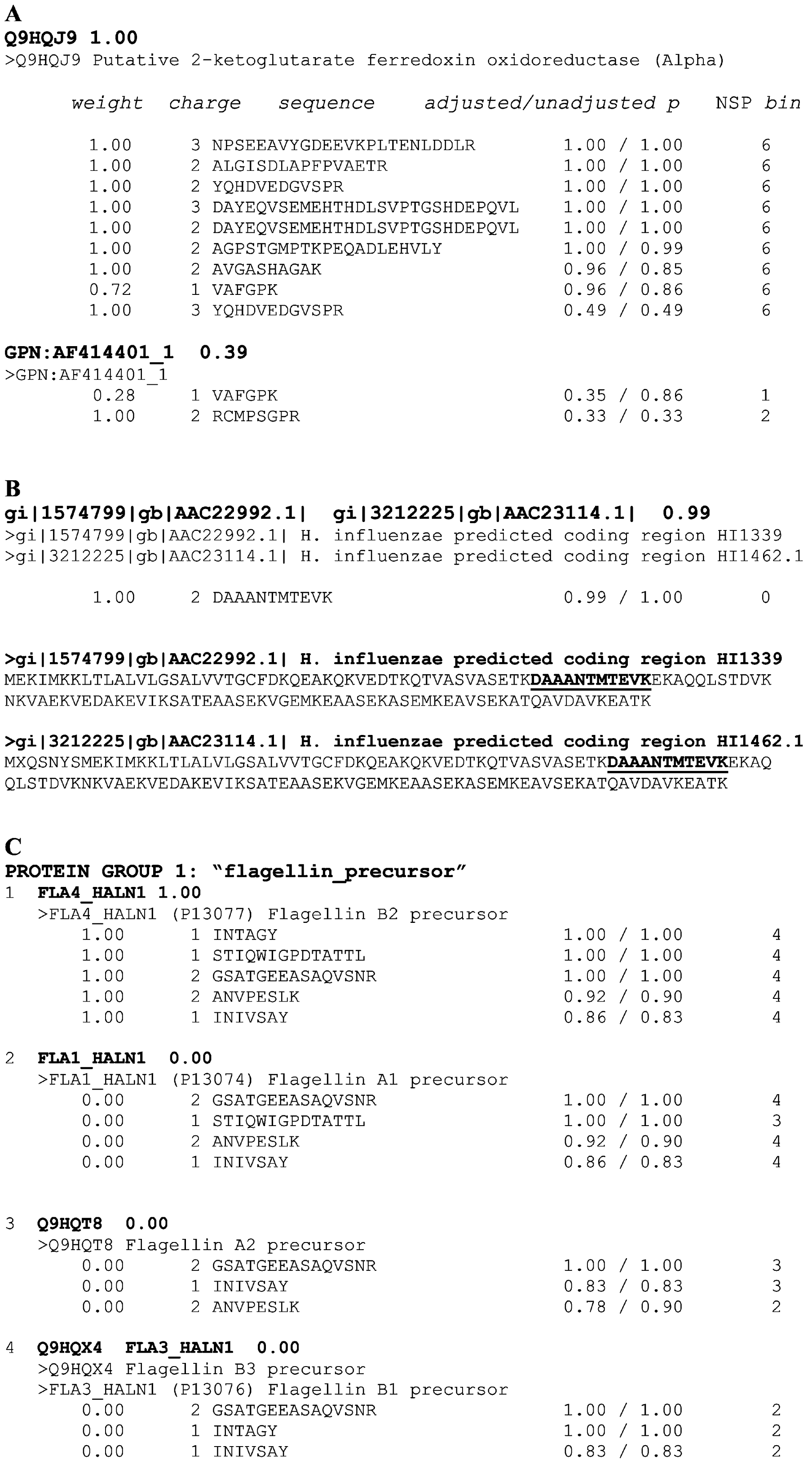
Data sets Hinf_Hum and Halo_Hum both contain a number of degenerate peptides, each corresponding to more than a single protein in the database, which can be used to evaluate how the model handles such cases. Figure 7A illustrates a case taken from the Halo_Human data set, where a peptide VAFGPK is shared between two unrelated proteins, Q9HQJ9 from the sample organism, *Halobacterium* (correct protein identification) and a human protein GPN:AF414401_1 (incorrect identification). The initial (unadjusted for NSP) probability that the peptide VAFGPK is a correct identification, given its database search scores, number of tryptic termini, and number of missed cleavages, is 0.86. Both proteins sharing this peptide have other nondegenerate peptides corresponding to them. Q9HQJ9, however, has a significantly larger number of peptides that correspond to it, while GPN:AF414401_1 has only one other peptide in the data set, RCMPSGPR, with initial probability 0.33, which is an incorrect assignment. Thus, peptide VAFGPK has many sibling peptides and a very high NSP value with respect to protein Q9HQJ9 (NSP value 14.66, NSP bin 6) and only one low probability sibling, and a much lower NSP value, with respect to protein GPN:AF414401_1 (NSP value 0.33, NSP bin 1). After adjustment for NSP, its probability is increased to 0.97, assuming it corresponds to protein Q9HQJ9, and decreased to 0.35, assuming it corresponds to GPN: AF414401_1, and its apportionments between the two proteins are estimated to be 0.72 and 0.28, respectively. Thus, the model correctly predicts that it is much more likely that peptide VAFGPK is present in the tryptic digest due to the presence in the original sample of protein Q9HQJ9 rather than GPN:AF414401_1. Since GPN:AF414401_1 does not have any other significant evidence, it is assigned a relatively low probability (0.39) of being present in the sample. In contrast, it would have been assigned a probability of 0.91 without adjustment of peptide probabilities to account for NSP and assuming a full contribution from degenerate peptide VAFGPK.

Another example, taken from the Hinf_Hum data set, is shown in Figure 7B. A peptide DAAANTMTEVK, identified with very high probability, is present in two protein sequence database entries corresponding to two predicted coding regions, HI1339 and HI1462.1. The only difference between these two entries is that HI1462.1 has an additional stretch of six amino acids at its N terminus (Figure 7B). No other peptide corresponding to either of these two proteins was found in the data set. Therefore, given the available mass spectrometric information, and without any additional knowledge such as the molecular weight of the sample protein, it is impossible to distinguish between these two protein entries and to determine which protein is more likely to be present in the sample. As a result, both entries, HI1339 and HI1462.1, are reported together as a single identification. This example illustrates an important point that a high-throughput peptide sequencing approach might not be sufficient to distinguish between proteins with a high degree of sequence similarity, such as splicing variants of the same gene product. In the particular case shown in Figure 7B, for example, to distinguish HI1462.1 from HI1339, one would need to look specifically for a peptide spanning the N terminal region of the former.

Figure 7C shows an even more complicated situation observed in the Halo_Hum data set. A total of five peptides were identified corresponding to a group of flagellin precursor proteins. In addition, there are no peptides corresponding to only one of the proteins in the group and not another. One of the proteins, FLA4_HALN1, corresponds to five peptides, while all other proteins in the group correspond to only a subset of the five identified peptides. In accordance with the Occam's razor approach implemented in the model, FLA4_HALN1 is considered the most probable candidate since it is sufficient to explain the presence of all five identified peptides in the tryptic digest of the sample. Thus, FLA4_HALN1 is assigned a probability close to 1 and is apportioned the full weights of its shared peptides, while all other proteins in the group are assigned 0 probability. Nevertheless, the presence of FLA4_HALN1 is not required to explain the observed data. For example, the observed peptides could have originated from the presence of both FLA_HALN1 and Q9HQX4 in the sample. The model therefore presents the entire set of similar flagellin precursor proteins as a group in order to assist the user in interpretation of the data. Note that situations of the kind discussed here, while quite rare in the *H. influenzae* or *Halobacterium* test data sets used to evaluate the performance of the model in this work, occur more frequently in data sets generated from higher eukariotic organisms. Preliminary results of applying this method to analyze the data generated in large-scale experiments performed on human raft cells are encouraging and demonstrate satisfactory grouping of similar proteins and apportioning of degenerate peptides among their corresponding proteins.[45]

General Utility of the Model. Probabilities computed by the model are accurate measures of confidence to accompany protein identifications and provide a standardized way of publishing large-scale proteomics data sets in the literature. For example, researchers can filter data below a low minimum probability threshold, such as 0.1, to remove the majority of incorrect results, and then publish the remaining protein list along with the computed probabilities that those identifications are correct. This allows users to have access to the most complete data set possible to interpret or further utilize at their discretion, as long as the reported protein probabilities are given full consideration. A researcher desiring only the most confident identifications, as is often the case for high-throughput experiments, can accept only those reported identifications with a high probability (for example, at least 0.9). On the other hand, a researcher interested in a particular protein has the opportunity to observe the evidence for its presence in the sample, no matter how slight. Inconclusive evidence might be sufficient justification for additional experiments to determine the validity of that identification. Published protein identifications accompanied by accurate probabilities would also provide maximal information to higher level computational analyses based on proteomics data, such as those concerned with the identification of protein−protein interactions, as long as they take the protein probabilities into account.

[45] von Haller, P. D. et al. Manuscript in preparation.



```
A
Q9HQJ9 1.00
>Q9HQJ9 Putative 2-ketoglutarate ferredoxin oxidoreductase (Alpha)

      weight  charge   sequence    adjusted/unadjusted p   NSP bin

        1.00      3 NPSEEAVYGDEEVKPLTENLDDLR        1.00 / 1.00       6
        1.00      2 ALGISDLAPFFPVAETR               1.00 / 1.00       6
        1.00      2 YQHDVEDGVSPR                    1.00 / 1.00       6
        1.00      3 DAYEQVSEMEHTHDLSVPTGSHDEPQVL    1.00 / 1.00       6
        1.00      2 DAYEQVSEMEHTHDLSVPTGSHDEPQVL    1.00 / 1.00       6
        1.00      2 AGPSTGMPTKPEQADLEHVLY           1.00 / 0.99       6
        1.00      2 AVGASHAGAK                      0.96 / 0.85       6
        0.72      1 VAFGPK                          0.96 / 0.86       6
        1.00      3 YQHDVEDGVSPR                    0.49 / 0.49       6

GPN:AF414401_1  0.39
>GPN:AF414401_1
        0.28      1 VAFGPK                          0.35 / 0.86       1
        1.00      2 RCMPSGPR                        0.33 / 0.33       2


B
gi|1574799|gb|AAC22992.1|  gi|3212225|gb|AAC23114.1|  0.99
>gi|1574799|gb|AAC22992.1| H. influenzae predicted coding region HI1339
>gi|3212225|gb|AAC23114.1| H. influenzae predicted coding region HI1462.1

        1.00      2 DAAANTMTEVK                     0.99 / 1.00       0


>gi|1574799|gb|AAC22992.1| H. influenzae predicted coding region HI1339
MEKIMKKLTLALVLGSALVVTGCFDKQEAKQKVEDTKQTVASVASETKDAAANTMTEVKEKAQQLSTDVK
NKVAEKVEDAKEVIKSATEAASEKVGEMKEAASEKASEMKEAVSEKATQAVDAVKEATK

>gi|3212225|gb|AAC23114.1| H. influenzae predicted coding region HI1462.1
MXQSNYSMEKIMKKLTLALVLGSALVVTGCFDKQEAKQKVEDTKQTVASVASETKDAAANTMTEVKEKAQ
QLSTDVKNKVAEKVEDAKEVIKSATEAASEKVGEMKEAASEKASEMKEAVSEKATQAVDAVKEATK


C
PROTEIN GROUP 1: "flagellin_precursor"
1  FLA4_HALN1 1.00
   >FLA4_HALN1 (P13077) Flagellin B2 precursor
       1.00      1 INTAGY                          1.00 / 1.00        4
       1.00      1 STIQWIGPDTATTL                  1.00 / 1.00        4
       1.00      2 GSATGEEASAQVSNR                 1.00 / 1.00        4
       1.00      2 ANVPESLK                        0.92 / 0.90        4
       1.00      1 INIVSAY                         0.86 / 0.83        4

2  FLA1_HALN1  0.00
   >FLA1_HALN1 (P13074) Flagellin A1 precursor
       0.00      2 GSATGEEASAQVSNR                 1.00 / 1.00        4
       0.00      1 STIQWIGPDTATTL                  1.00 / 1.00        3
       0.00      2 ANVPESLK                        0.92 / 0.90        4
       0.00      1 INIVSAY                         0.86 / 0.83        4


3  Q9HQT8  0.00
   >Q9HQT8 Flagellin A2 precursor
       0.00      2 GSATGEEASAQVSNR                 1.00 / 1.00        3
       0.00      1 INIVSAY                         0.83 / 0.83        3
       0.00      2 ANVPESLK                        0.78 / 0.90        2

4  Q9HQX4  FLA3_HALN1  0.00
   >Q9HQX4 Flagellin B3 precursor
   >FLA3_HALN1 (P13076) Flagellin B1 precursor
       0.00      2 GSATGEEASAQVSNR                 1.00 / 1.00        2
       0.00      1 INTAGY                          1.00 / 1.00        2
       0.00      1 INIVSAY                         0.83 / 0.83        2
```

Figure 7. Examples of model results with degenerate peptides (see text for details). (A) In the Hinf_Hum data set, identified peptide VAFGPK is shared between an unrelated correct (Q9HQJ9) and incorrect (GPN:AF414401_1) protein. The model correctly determines that it is much more likely that peptide VAFGPK is present in the tryptic digest due to the presence in the original sample of protein Q9HQJ9 rather than GPN:AF414401_1. (B) In the Hinf_Hum data set, two protein database entries, HI1339 and HI1462.1, are not distinguishable on the basis of the single observed peptide identification and are reported together as a single identification. (C) In the Halo_Hum data set, a total of five peptides are identified corresponding to a group of flagellin precursor proteins, none of which has any nondegenerate evidence. One of the proteins, FLA4_HALN1, contains all five peptides, while all other proteins in the group contain only a subset of the five identified peptides. FLA4_HALN1 is therefore the most probable candidate since its presence in the sample is sufficient to explain the presence of all identified peptides in the tryptic digest of the sample. These proteins are presented by the model as a group ("flagellin_precursor").

The model can predict the number of correct protein identifications in data sets, as well as the error rates resulting from filtering the data with any minimum probability threshold. This information could serve as objective criteria by which related data sets are compared, regardless of differences in the experimental or computational methods used to generate the data. For example, one can compare unfiltered protein identification data sets using the total number of correct protein identifications predicted by the model. In addition, one can compare filtered data sets objectively by specifying a uniform error rate and applying to each data set the corresponding minimum probability threshold. The model predicted error rate and the number of correct protein identifications can similarly be used to compare the performance of different computational methods, such as different database search tools, and different experimental protocols.

At present, this model has been applied to large-scale data sets of peptide assignments produced by MS/MS database search tools. It should be stressed that it can be applied to data sets of peptide assignments produced by *any* database search tool, or by a combination of several such tools, as long as each peptide assignment is accompanied by an accurate probability that it is correct. Peptide probabilities can be considered accurate (in the context of the complete data set for which they are calculated), if upon selection of all peptides in the data set having any given computed probability, the corresponding proportion of them is correct. Suitable peptide probabilities for any database search tool can be obtained using the software PeptideProphet or a similar implementation of the peptide statistical model described in ref 25. Note that the probabilities that peptide assignments are correct can be computed independently for disparate data sets (e.g., spectra collected from different samples or using different mass spectrometers) and then combined prior to analysis. Finally, the statistical model can in principle be applied to peptides assigned to MS/MS spectra using computational methods not relying exclusively on the database search, as well as to peptides identified using experimental methods other than MS/MS sequencing.

Future Work. One possible improvement of the model is related to the definition of NSP, given in eq 4. More accurate protein probabilities might be achieved by renormalizing NSP to account for the differences in the expected number of peptides among different proteins. In addition, an empirical factor reflecting common knowledge about the protein digestion process can be introduced in eq 4.[46] Various ways of renormalizing NSP are being investigated.

An interesting future pursuit is to explore various ways to incorporate prior knowledge about the biological system in general, and the samples analyzed by mass spectrometry in particular, when such information is available. For example, assume that it is known that the given sample consists of predominantly nuclear fraction proteins. If then a peptide is observed that is shared between two proteins, A and B, with protein A being more likely to be present in a nuclear fraction than B, it could then be inferred more likely that the peptide is present in the peptide mixture due to the presence of protein A in the original sample rather than protein B. Thus, in this particular example, the prior knowledge regarding cellular localization could be useful for resolving some cases of degenerate peptides. In another example, corresponding to the situation shown in Figure 4, assume that it is determined, e.g., via the ICAT labeling approach, that both peptides, 1 and 2, are present in the sample at significantly different relative abundances with respect to a control sample. This, in turn, would indicate that *both* proteins, A and B, are likely to be present but at different relative abundance levels. In general, prior information could be any knowledge available in the literature or obtained from other kinds of measurements performed on the same systems in the course the study. In fact, all pieces of information mentioned above are routinely taken into consideration by biologists analyzing and validating protein identifications obtained in their experiments.

CONCLUSIONS

The described method for computing probabilities that proteins are present in a sample on the basis of peptides assigned to MS/MS spectra acquired from a proteolytic digest of the sample enables high-throughput analysis of large-scale proteomics experiments. It produces accurate probabilities that proteins are present in the sample, with high power to discriminate correct from incorrect protein identifications. It is fully automated and fast and does not rely on subjective manual validation. The method allows filtering of large-scale data sets with predictable sensitivity and false positive identification error rates. It presents results in an organized manner by collapsing redundant protein sequence database entries into single identifications and by grouping together proteins that are not distinguishable on the basis of peptides assigned to MS/MS spectra. It provides a new standard for publishing large-scale proteomics data sets in the literature and enables the comparison of results from different research groups, obtained using different experimental protocols, different mass spectrometers, and even different MS/MS database search tools. Resulting lists of protein identifications along with their computed probabilities can also serve as useful inputs to computational tools being developed that rely on the data generated in high-throughput proteomics studies, such as those concerned with the analysis of protein−protein interaction networks and metabolic pathway reconstruction.

The software ProteinProphet implementing the statistical model described in this work will be available to the public at http://systemsbiology.org/research/software.html.

ACKNOWLEDGMENT

This work was funded in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, under Contract N01-HV-28179, from the National Cancer Institute, NIH, Grant 1-R33-CA93302, and by sponsored research agreement from Oxford GlycoSciences to R.A., and by the DOE's Offices of Biological and Environmental Research and Advanced Scientific Computing Research *Genomes to Life* Grant DE-FG08-01ER63218 to E.K. We thank Anne Claude Gingras, Jimmy Eng, Dan Martin, Xiao-jun Li, Victor Ng, Jeff Ranish, Andy Siegel, and Julian Watts for helpful discussions, and Bill Noble for critical reading of the manuscript. We also appreciate contributions from David Goodlett, Sergey Stolyar, Sam Purvine, and Eugene Yi regarding the experimental data sets used in this study.

[46] Zhang, W.; Chait, B. T. *Anal. Chem.* **2000**, *72*, 2482.

SUPPORTING INFORMATION AVAILABLE

Coefficients of the derived discriminant function used to compute probabilities that SEQUEST search results of singly charged precursor ion spectra are correct, as well as data demonstrating improved discriminating power of peptide probabilities computed with adjustment for NSP and accuracy of computed peptide probabilities for the Halo_Hum and Hinf_Hum data sets. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review February 7, 2003. Accepted June 5, 2003.

AC0341261