ADVANCEMENT IN PROTEIN INFERENCE FROM SHOTGUN PROTEOMICS USING PEPTIDE DETECTABILITY

PEDRO ALVES,[1] RANDY J. ARNOLD,[2] MILOS V. NOVOTNY,[2] PREDRAG RADIVOJAC,[1] JAMES P. REILLY,[2] HAIXU TANG[1,3*]

1) School of Informatics, 2) Department of Chemistry,
3) Center for Genomics and Bioinformatics, Department of Biology
Indiana University, Bloomington, U.S.A.

A major challenge in shotgun proteomics has been the assignment of identified peptides to the proteins from which they originate, referred to as the *protein inference problem*. Redundant and homologous protein sequences present a challenge in being correctly identified, as a set of peptides may in many cases represent multiple proteins. One simple solution to this problem is the assignment of the smallest number of proteins that explains the identified peptides. However, it is not certain that a natural system should be accurately represented using this minimalist approach. In this paper, we propose a reformulation of the protein inference problem by utilizing the recently introduced concept of peptide detectability. We also propose a heuristic algorithm to solve this problem and evaluate its performance on synthetic and real proteomics data. In comparison to a greedy implementation of the minimum protein set algorithm, our solution that incorporates peptide detectability performs favorably.

1. Introduction

Shotgun proteomics refers to the use of bottom-up proteomics techniques in which the protein content in a biological sample mixture is digested prior to separation and mass spectrometry analysis.[1-3] Typically, liquid chromatography (LC) is coupled with tandem mass spectrometry (MS/MS) resulting in high-throughput peptide analysis. The MS/MS spectra are searched against a protein database to identify peptides in the sample. Currently, Sequest[4] and Mascot[5] are the most frequently used computer programs for conducting peptide identification, both comparing experimental MS/MS spectra with *in silico* spectra generated from the peptide sequences in a database. Compared to top-down proteomics techniques, shotgun proteomics avoids the modest separation efficiency and poor mass spectral sensitivity associated with intact protein analysis, but it also encounters a new problem in data analysis, that of determining the set of proteins present in the sample based on the peptide identification results. At a

* To whom all correspondence should be addressed; Email: hatang@indiana.edu.

first glance, this problem seems trivial. It may be concluded that a protein is present in the sample, if and only if at least one of its peptides is identified. This conclusion is true, however, only when each identified peptide is *unique*, i.e. when it belongs to only one protein. If some peptides are *degenerate*,[6] i.e. shared by two or more proteins in the database, determining which of these proteins exist in the sample has multiple possible solutions. Indeed, tryptic peptides are frequently degenerate, especially for the proteome samples of vertebrates, which, due to recent gene duplications, often have a large number of paralogs. In addition, alternative splicing in higher eukaryotes results in many identical protein subsequences. The following example illustrates the extent of peptide degeneracy in a real proteomics experiment. Of the 693 identified peptides from a real rat sample used in this study (see sections 3-4 for details), 296 were unique and 397 were degenerate, when searched against the full proteome of *R. norvegicus*. These peptides can be assigned to a total of 805 proteins, of which only 149 proteins could be assigned based on the 296 unique peptides.

Nesvizhskii and colleagues first formalized this challenge in shotgun proteomics data analysis. They formulated the *protein inference problem* and proposed a solution as the minimum number of proteins containing the set of identified peptides.[6, 7] Other methods assign the *unique* peptides first, and then use statistical methods[6] to assign the degenerate peptides based on the likelihood of each putative protein already identified. As a result, if two proteins share some common tryptic peptides, the presence of each protein can be decided using this method only if there exists at least one identified unique peptide in one of the proteins. The degenerate peptides will be most likely assigned to the longer protein, because the shorter proteins may not contain any unique peptide (e.g. see Fig. 2 in reference 7).

In this paper, we revisit the protein inference problem based on the recently proposed concept of *peptide detectability*.[8] The detectability of a peptide is defined as the probability of observing it in a standard proteomics experiment. We proposed that detectability is an intrinsic property of a peptide, completely determined by its sequence and its parent protein. We also showed that the peptide detectability can be estimated from its parent protein's primary structure using a machine learning approach.[8] The introduction of peptide detectability provides a new approach to protein inference, in which not only identified peptides but also those that are missed (not identified) are important for the overall outcome. Figure 1 illustrates the advantage of the new idea. Assume *A* and *B* are two proteins sharing 3 degenerate tryptic peptides (*a*, *b*, and *c*, shaded). Each protein in Fig. 1 also has unique tryptic peptides (*d*, *e*, and *f*, *g*, *h*, *i* respectively, white). According to the original formulation of the protein inference problem, the identities of *A* and *B* cannot be determined since the only identified peptides

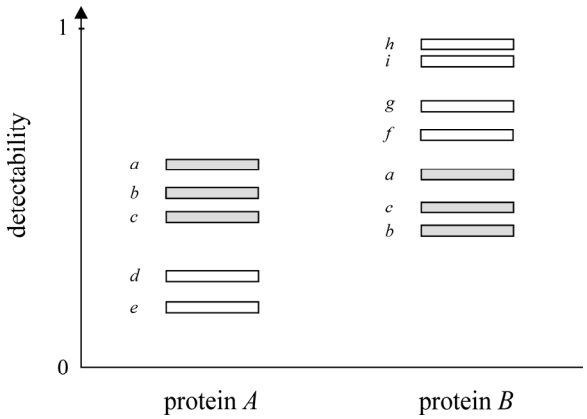

Figure 1. Detectability plot of a hypothetical protein *A*, broken up into tryptic peptides *a*-*e*, and protein *B*, containing peptides *a*-*c* and *f*-*i*. Assume that peptides *a*-*c* are identified by the peptide identification software (shaded). Peptides in each protein are sorted according to their detectability. The example shows the intuition for tie breaking in the proposed protein inference problem. Peptides *a*-*c* are more likely to be observed in protein *A* than *d*-*e*, while they are less likely to be observed than peptides *f*-*i* in protein *B*. Thus, protein *A* is a more likely to be present in the sample than *B*. Note that the detectability for the same peptide within different proteins is not necessarily identical, due to the influence of neighboring regions in its parent proteins.

are degenerate.[7] However, if all the tryptic peptides are ranked in each protein according to their detectabilities (Fig. 1), we may infer that protein *A* is more likely to be present in the sample than protein *B*. This is because if *B* is present we would have probably observed peptides *f*-*i* along with peptides *a*-*c*, which all have lower detectabilities than either *f*, *g*, *h*, or *i*. On the other hand, if protein *A* is present, we may still miss peptides *d* and *e*, which have lower detectabilities than peptides *a*-*c*, especially if *A* is at relatively low abundance.[8] In summary, peptide detectability and its correlation with protein abundance provides a means of inferring the likelihood of identifying a peptide relative to all other peptides in the same parent protein. This idea can then be used to distinguish between proteins that share tryptic peptides based on a probabilistic framework.

Based on this simple principle, we propose a reformulation of the protein inference problem so as to exploit the information about computed peptide detectabilities. We also propose a tractable heuristic algorithm to solve this problem. The results of our study show that this algorithm produces reliable and less ambiguous protein identities. These encouraging results demonstrate that peptide detectability can be useful for not only label-free protein quantification, but also for protein identification that is based on identified peptides.[8, 9]

2. Problem Formulation

Consider a set of proteins $\mathcal{P} = \{P_1, P_2, \ldots, P_N\}$ such that each protein P_j consists of a set of tryptic peptides $\{p_j^i\}$, $i = 1, 2, \ldots, n_j$, where n_j is the number of peptides in $\{p_j^i\}$. Suppose that $\mathcal{F} = \{f_1, f_2, \ldots, f_M\}$ is the set of peptides identified by some database search tool and that $\mathcal{F} \subseteq \cup\{p_j^i\}$. Finally, assume each peptide p_j^i has a computed detectability $D(p_j^i)$, for $j = 1, 2, \ldots, N$, and $i = 1, 2, \ldots, n_j$. We use $\mathcal{D}$ to denote the set of all detectabilities $D(p_j^i)$, for each i and j.

The goal of a protein inference algorithm is to assign every peptide from $\mathcal{F}$ to a subset of proteins from $\mathcal{P}$ which are actually present in the sample. We call this assignment the *correct peptide assignment*. However, because in a real proteomics experiment the identity of the proteins in the sample is unknown, it is difficult to formulate the fitness function that equates optimal and correct solutions. Thus, the protein inference problem can be redefined to find an algorithm and a fitness function which result in the peptide-to-protein assignments that are *most probable*, given that the detectability for each peptide is accurately computed. In a practical setting, the algorithm's optimality can be further traded for its robustness and tractability.

If all peptides in $\mathcal{F}$ are required to be assigned to at least one protein, the choice of the likelihood function does not affect the assignment of unique (non-degenerate) peptides in $\cup\{p_j^i\}$. On the other hand, the tie resolution for degenerate peptides will depend on all the other peptides that can be assigned to their parent proteins, and their detectabilities. In order to formalize our approach we proceed with the following definitions.

Definition 1. Suppose that the peptide-to-protein assignment is known. A peptide $p_j^i \in \{p_j^i\}$ is considered *assigned* to P_j if and only if $p_j^i \in \mathcal{F}$ and $D(p_j^i) \geq M_j$. Then, $M_j \in \mathcal{D}$ is called the *Minimum Detectability of Assigned Peptides* (MDAP) of protein P_j.

Definition 2. A set of MDAPs $\{M_j\}_{j = 1, 2, \ldots, N}$ is *acceptable* if for each $f \in \mathcal{F}$, there exists P_j, such that $D(f) \geq M_j$. Thus, any *acceptable* MDAP set will result in an assignment of identified peptides that guarantees that every identified peptide is assigned to at least one protein.

Definition 3. A peptide p_j^i is *missed* if $p_j^i \notin \mathcal{F}$ and $D(p_j^i) \geq M_j$.

Note that, due to the connection between peptide detectability and protein amount in the sample, peptides whose detectabilities are below M_j are not considered missed. We can now formulate the protein inference problem as follows.

Minimum missed peptide problem. Given N proteins, each consisting of n_j tryptic peptides, and a set of identified peptides $\mathcal{F}$, find an acceptable set of MDAPs, $\{M_j\}_{j = 1, 2, \ldots, N}$, which result in a minimum number of missed peptides.

If a protein does not exist in the sample, the MDAP M_j needs to be assigned a value greater than the maximum detectability observed in P_j. If protein P_j is not present in the sample, we set M_j to a maximum MDAP (= ∞). Hence, only proteins whose $M_j \leq 1$ are considered identified. Note that in nearly all practical cases the maximum MDAP can be set to 1, except when there is a peptide in $\cup \{p_j^i\}$ whose $D(p_j^i) = 1$. The relationship between the minimum missed peptide problem and the original minimum protein set problem becomes evident in the following theorem.

Theorem 1. Minimum missed peptide problem is NP-hard.

Proof: The minimum missed peptide problem can be reduced to the set-covering problem[10] by setting $D(p_j^i) = 0$ for each i, j and adding a non-existing peptide with detectability of 1 to each protein. Minimizing the number of missed peptides now minimizes the number of covering subsets (proteins) in the solution set.

□

3. Materials and Methods

3.1. *Data*

The data used in this paper were obtained from three different sources. Our first two datasets were generated using mixtures of model proteins. Therefore, we know the proteins in these two samples. The first set was generated as a standard protein mixture consisting of 12 model proteins and 23 model peptides mixed at similar concentrations from 73 to 713 nM for proteins and from 50 to 1800 nM for peptides.[11] This data set was made available to us by the authors.

The second data set from a mixture of twelve standard proteins[8] was prepared at 1 μM of final digestion solution for each protein except human hemoglobin which is at 2 μM (mixture B), combined with buffer, reduced, alkylated, and digested overnight with trypsin. Peptides were separated by nano-flow reversed-phase liquid chromatography gradient and analyzed by mass spectrometry and tandem mass spectrometry in a Thermo Electron (San Jose, CA) LTQ linear ion trap mass spectrometer.

The third sample was generated using a complex proteome sample from *R. norvegicus*. Rat brain hippocampus samples were homogenized and separated by sedimentation in a centrifuge to produce four fractions enriched in nuclei, mitochondria, microsomes (remaining organelles), and the cytosol. Each subcellular fraction was subjected to proteolytic digestion with trypsin and analyzed by reversed-phase capillary LC tandem mass spectrometry using a 3-D ion trap (ThermoFinnigan LCQ Deca XP). Searches versus either the Swiss-Prot or the

IPI rat database were performed for fully tryptic peptides using Mascot[5] with a minimum score of 40 and allowing for N-terminal protein acetylation and methionine oxidation.

3.2. *Prediction of peptide detectability*

As mentioned above, the probability that a peptide will be identified in a standardized proteomics experiment is referred to as the peptide detectability.[8] Using machine learning approaches we previously provided evidence that peptide detectability can be predicted solely from the amino acid sequence of its parent protein. We constructed a set of 175 features describing the peptide sequence itself as well as the regions up or downstream from the peptide. An ensemble of neural networks was then trained and evaluated. We estimated its balanced-sample accuracy at about 70% across training and test sets obtained from several independent proteomics studies. The usefulness of the learned peptide detectabilities was demonstrated on the problem of label-free protein quantification where the detectability of a peptide showed negative correlation with the abundance of its parent protein.

3.3. *Solving minimum missed peptide problem*

We propose a simple greedy algorithm to solve the minimum missed peptide problem. It assigns identified peptides to proteins in the order of their detectabilities and does not change the peptide assignments once they are made. The algorithm assigns the peptide with lowest detectability first (denoted as Lowest-Detectability First Algorithm, LDFA). The pseudocode for the LDFA is presented in Fig. 2. We assume the detectabilities of a single peptide across different parent proteins are close enough not to affect the relative order of each such peptide in its parent protein if a representative detectability is selected. Thus, all identified peptides can be sorted consistently based on their detectabilities.

For comparison with LDFA, we also implement a greedy solution to the minimum protein set algorithm (GMPSA), which can be formulated as a set-covering problem[10] with very little modification.

4. Results

We compared the performance of the LDFA and GMPSA. First, we used identified peptides from a synthetic sample mixture B,[8] and Swiss-Prot as a reference database to conduct a controlled protein inference experiment. The advantage of this evaluation for quantifying the performance of the algorithm is that all proteins present in the sample are known. The sample mixture B contained 12 proteins corresponding to the 93 peptides identified in the experiment.

Algorithm. Lowest-detectability first algorithm (LDFA)

Assign all unique peptides in $\mathcal{F}$ and remove them from $\mathcal{F}$
$M_j = \infty$ for all j's
while $\mathcal{F} \neq \varnothing$
    Choose $f \in \mathcal{F}$ with lowest detectability
    for each protein i containing f
        Compute the number of missed peptides, assuming
        $M_i = D(f)$
    Select protein j with the minimum number of missed peptides
    Set $M_j = D(f)$
    Remove from $\mathcal{F}$ all peptides from protein j

Figure 2. Pseudocode for the LDFA solution to the minimum missed peptide problem.

Out of 176,470 proteins from Swiss-Prot, 494 proteins (including the 12 proteins from the mixture) were identified as containing at least one identified peptide. The LDFA identified 12 proteins in the sample, 11 correctly. Of the 11 proteins that were correctly assigned, in only one instance could the algorithm not distinguish between the correct protein and one of its close homologs. We refer to this situation as a *tie*. Each tie is resolved by a random selection.

The same data was tested using the GMPSA which simply tries to explain the identified peptides with the smallest possible number of proteins. GMPSA also identified 12 proteins as the total number of proteins in the sample, however, it suffered in accuracy. For 5 out of the 12 proteins, the GMPSA could not distinguish between the correct proteins and their homologs. Since in each step, the GMPSA considers only the number of the identified peptides per protein it is much more likely to encounter ties than the LDFA. As shown in Fig. 1, the GMPSA does not have a means of differentiating between proteins containing no unique identified peptides and the same number of degenerate peptides. In practice, these result in ties involving more homologs than the LDFA, thus reduce the chance of selecting the correct protein. An example of such a tie involves protein HBB_HUMAN. LDFA found two possible solutions (HBB_HUMAN and HBB_GORGO), resulting in a 50% chance of a correct selection. On the other hand, the GMPSA selected between four different proteins (HBB_HUMAN, HBB_HAPGR, HBB_HYLLA and HBB_PANPO) resulting in 25% chance of a correct prediction. Furthermore, the smaller average number of proteins per tie encountered by LDFA is advantageous for reporting results of identification. To avoid information leak in calculating peptide detectabilities, the training set for the predictor was constructed from a different synthetic dataset.[11]

The one protein that was not identified correctly by the LDFA, bovine RNase A, was assigned to a close homolog from one of 7 organisms (69.4% average sequence identity) chosen at random. This assignment was made with a single identified peptide. Furthermore, the sequence for bovine RNase A in the Swiss-Prot database includes the 26-amino acid signal peptide that is not actually present in the sample. Since LDFA takes into consideration the detectabilities of both identified and unidentified peptides, the presence of the signal peptide in the database hinders the assignment of bovine RNase A. After the signal peptide is removed, the sequence identity compared to all seven sequences that match the identified peptide is 84.0%. In comparison, the GMPSA randomly selects among 20 proteins from Swiss-Prot sharing the identified peptide.

Another experiment was performed on a biological sample from *R. norvegicus*, in which the correct proteins were not known. The identified peptides in the sample (693 in total) were searched against an IPI (http://ncbi.nlm.nih.gov) database and were found in 805 proteins. These are the proteins that may potentially be present in the sample. Table 1 shows the distribution of these peptides contained by different numbers of proteins. In this experiment, about 60% identified peptides (397 out of 693) are degenerate peptides, i.e. contained by two or more proteins. The two algorithms described above, LDFA and GMPSA, were run on this set.

Table 1. Distribution of identified peptides contained by different number of proteins in a *R. norvegicus* proteome analysis.

No. proteins	1	2-5	6-10	11-20	>20
No. peptides	296	330	43	16	8

Mascot had originally assigned 301 proteins in this sample, LDFA assigned 275 proteins and GMPSA assigned 247 proteins. Taking into consideration all unique peptides from the rat sample only 149 proteins could be assigned by at least one unique peptide. Thus, any other protein to be assigned by any of the three methods would have to rely solely on degenerate peptides. Due to the prevalence of ties, GMPSA was run 30 times. Only 153 proteins were consistently assigned in all runs. Out of 430 proteins assigned over all GMPSA runs, 229 were assigned less than 50% of the time.

Since the correct proteins in this sample were not known, the accuracy of the LDFA and GMPSA could not be quantified as on the synthetic data. Instead, a different approach was taken where protein distinguishability was measured in this experiment. Figure 3 shows, in grey, all pairs of 805 identified proteins that shared at least one identified peptide. The y-axis corresponds to the percentage of sequence identity, while the x-axis represents the length of one of the proteins

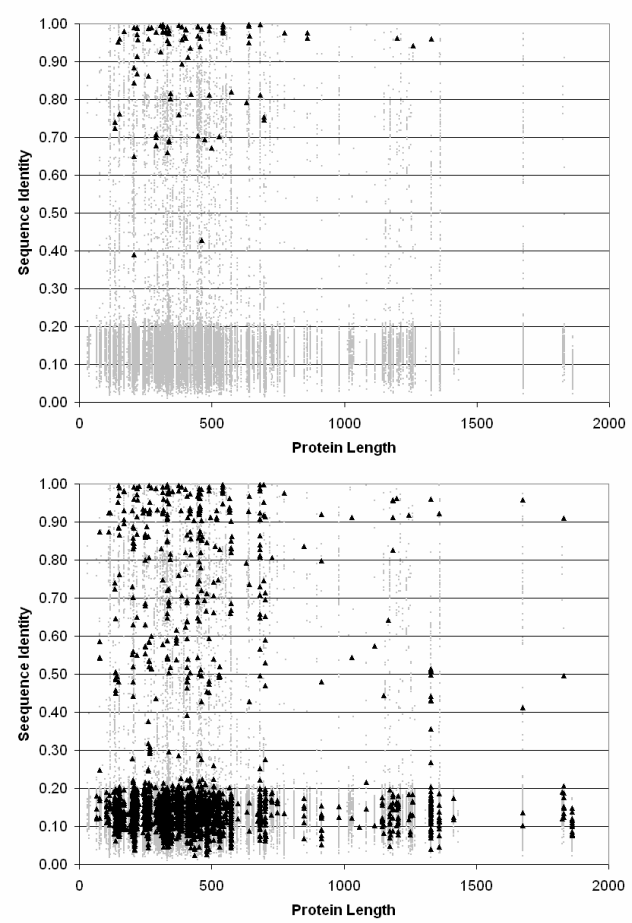

Figure 3. A pairwise comparison of all proteins in IPI rat database in which proteins share at least one identified peptide. The grey dots indicate all pairs while the black triangles indicate pairs where the algorithm made a random selection between the two proteins for a) LDFA and b) GMPSA.

in the pair. Figure 3a shows, in black, all pairs of proteins that share at least one identified peptide and that the LDFA could not distinguish. This means that at one point during the execution LDFA had to randomly select between those two proteins and that at the completion of the algorithm one of the proteins is not present in the final solution. Figure 3b shows the equivalent plot for the GMPSA. In a single run of each algorithm, there were 94 indistinguishable pairs for the LDFA and 2,346 indistinguishable pairs for the GMPSA. Interestingly,

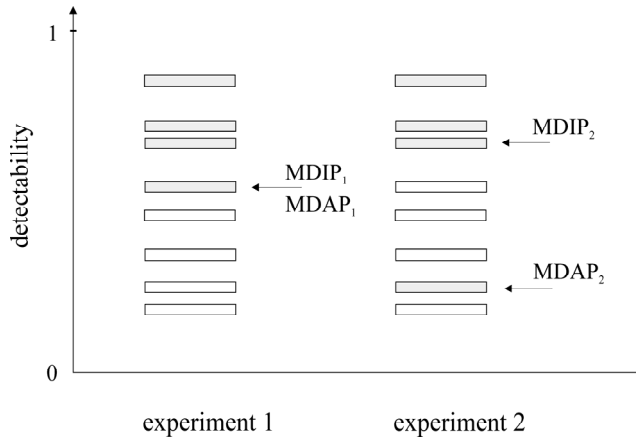

Figure 4. Detectability plot of a hypothetical protein consisting of 8 tryptic peptides from two shotgun proteomics experiments. Peptides that are identified are represented in grey. In experiment 1, MDIP is obtained as the detectability of an identified peptide that maximizes true positive (100%) and true negative (100%) rates. In experiment 2, the maximum true positive rate is 75%, while true negative rate is 100%.

the total number of proteins that were excluded from the final solution at random was 69 and 188 for the LDFA and GMPSA, respectively.

5. Discussion

In the previous study we have defined the Minimum acceptable Detectability of Identified Peptides (MDIP) as the detectability of an identified peptide that maximizes the average of the true positive and true negative rates for an identified protein. We also showed that MDIP of a protein is correlated with its abundance in the sample. The relationship between MDIP and MDAP is shown in Fig. 4 where the identified and non-identified peptides are shown for the same protein under two different experiments. While MDAP is always the lowest detectability of an identified peptide in a protein, MDIP is influenced by non-identified peptides as well. Ideally, as in the left part of Fig. 4, peptides are consecutively identified according to their decreasing detectabilities (starting from the top one), thus giving MDIP = MDAP. Non-identified peptides in the right part of Fig. 4 allow discrepancy between these two quantities which we believe will be useful for the advancement of label-free protein quantification.

One challenge in correctly interpreting shotgun proteomics data involves assigning identified peptides to the proteins from which they originate.[1, 3, 7, 12-15]

When the same peptide can be assigned to multiple proteins, this task – referred to as the protein inference problem – is non-trivial. Here we address this problem by utilizing the concept of peptide detectability – the probability that a peptide will be identified in a shotgun proteomics experiment based on inherent properties of the peptide and its surroundings within a protein. Previous work has shown that the rules governing peptide detectability can be assigned using a machine learning approach and that a peptide's detectability depends on its source protein concentration.[8] In cases where a peptide sequence is found in multiple protein sequences, knowledge of the detectabilities of both the identified peptides (similar sequences in the multiple proteins) and the unidentified peptides (some of which will differ in the multiple proteins) can be used to discern between assignments that would not otherwise be distinguishable.

The results shown here for 693 peptides identified from a rat brain sample indicate that 247 proteins can be assigned using a greedy algorithm for the minimum protein coverage formulation, but 94 (38%) of these are selected randomly. When peptide detectability is incorporated into the assignment algorithm, 275 proteins are assigned and only 51 (19%) of these are ambiguous. While the accuracy of this approach is difficult to test on a real proteomics data set, it is clear that the ability to distinguish potential peptide-to-protein assignments offers a significant advance in addressing the protein inference problem.

In a typical shotgun proteomics experiment, less than 10% identified tryptic peptides contain missed cleavages. Currently, we are not able to predict the detectabilities of these peptides because of the lack of training data. As a result, missed-cleavage peptides are neglected in protein inference even if they are identified. We aim to incorporate this prediction in the future. In this study, the identified peptides are determined based on a threshold of Mascot score 40, consistent with the condition used to generate the identified peptides in the dataset for training the detectability predictor.[8] If a different threshold is used, the predicted detectability may be different. The effects of threshold selection and other conditions used in peptide identification on the detectability prediction and protein inference will be explored in the future.

Acknowledgements

The authors wish to acknowledge the Office of the Vice President for Research for a Faculty Research Support Grant to RJA, PR, & HT. RJA and MVN acknowledge support from the 21st Century Fund (State of Indiana). JPR wishes to acknowledge support from NSF grant CHE-0518234.

References

1. Aebersold, R. & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature* **422**, 198-207.
2. McDonald, W. H. & Yates, J. R. r. (2003). Shotgun proteomics: integrating technologies to answer biological questions. *Curr Opin Mol Ther.* **5**, 302-309.
3. Kislinger, T. & Emili, A. (2003). Going global: protein expression profiling using shotgun mass spectrometry. *Curr Opin Mol Ther.* **5**, 285-293.
4. Yates, J. R., Eng, J. K., McCormack, A. L. & Schieltz, D. (1995). Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem* **67**, 1426-1436.
5. Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551-67.
6. Nesvizhskii, A. I., Keller, A., Kolker, E. & Aebersold, R. (2003). A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* **75**, 4646-4658. .
7. Nesvizhskii, A. I. & Aebersold, R. (2005). Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* **4**, 1419-1440.
8. Tang, H., Arnold, R. J., Alves, P., Xun, Z., Clemmer, D. E., Novotny, M. V., Reilly, J. P. & Radivojac, P. (2006). A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinformatics* **22**, (in press).
9. Gao, J., Opiteck, G. J., Friedrichs, M., Dongre, A. R. & Hefta, S. A. (2003). Changes in the protein expression of yeast as a function of carbon source. *J. Proteome Res.*, 643–649.
10. Cormen, T. H., Leiserson, C. E., Rivest, R. L. & Stein, C. (2001). *Introduction to algorithms*. 2nd edit, MIT Press, Cambridge, MA, U.S.A.
11. Purvine, S., Picone, A. F. & Kolker, E. (2004). Standard mixtures for proteome studies. *Omics* **8**, 79-92.
12. Carr, S., Aebersold, R., Baldwin, M., Burlingame, A., Clauser, K. & Nesvizhskii, A. (2004). The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data. *Mol Cell Proteomics* **3**, 531-3.
13. Rappsilber, J. & Mann, M. (2002). What does it mean to identify a protein in proteomics? *Trends Biochem Sci* **27**, 74-8.
14. 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. (2004). Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal Chem* **76**, 3556-68.
15. Yang, X., Dondeti, V., Dezube, R., Maynard, D. M., Geer, L. Y., Epstein, J., Chen, X., Markey, S. P. & Kowalak, J. A. (2004). DBParser: web-based software for shotgun proteomic data analyses. *J Proteome Res* **3**, 1002-8.