World Academy of Science, Engineering and Technology
International Journal of Biotechnology and Bioengineering
Vol:4, No:10, 2010

An Information Theoretic Approach to Rescoring Peptides Produced by De Novo Peptide Sequencing

John R. Rose, James P. Cleveland, Alvin Fox

Abstract—Tandem mass spectrometry (MS/MS) is the engine driving high-throughput protein identification. Protein mixtures possibly representing thousands of proteins from multiple species are treated with proteolytic enzymes, cutting the proteins into smaller peptides that are then analyzed generating MS/MS spectra. The task of determining the identity of the peptide from its spectrum is currently the weak point in the process. Current approaches to de novo sequencing are able to compute candidate peptides efficiently. The problem lies in the limitations of current scoring functions. In this paper we introduce the concept of proteome signature. By examining proteins and compiling proteome signatures (amino acid usage) it is possible to characterize likely combinations of amino acids and better distinguish between candidate peptides. Our results strongly support the hypothesis that a scoring function that considers amino acid usage patterns is better able to distinguish between candidate peptides. This in turn leads to higher accuracy in peptide prediction.

Keywords—Tandem mass spectrometry, proteomics, scoring, peptide, de novo, mutual information

I. Introduction

MS/MS is essential for high-throughput protein identification. The samples from which the data are derived may contain complex mixtures of numerous proteins. Moreover, in samples taken from the environment, the proteins may be extracted from multiple types of organisms. Mixtures are most commonly analyzed by on-line liquid chromatography-electrospray (ESI) MS/MS or off-line 2D gel electrophoresis followed by MALDI (matrix assisted laser desorption/time of flight) MS/MS. Regardless of how the mixtures are analyzed, proteins are treated with proteolytic enzymes to cut the proteins into smaller peptides of size that is manageable by MS/MS analysis. Peptide analysis by MS/MS generates product ion spectra [1]–[3]. The task of determining the identity of the peptide from its spectrum is currently the weak point in the process.

There are broadly three approaches to computer assisted peptide identification that use MS/MS data: database methods [4]–[7], de novo sequencing [8]–[13] and tagging [14], [15], which is a combination of de novo sequencing and database lookup. The database search methods are the workhorse for peptide identification in industry. There are commercial systems based on database search algorithms such as MASCOT and SEQUEST and are indicative of the maturity of this approach. The experimental spectrum is treated as a fingerprint and is compared with theoretical spectra computed for peptides in the database. The proteomic sequence database is largely generated from determination of a whole genome, by high-throughput DNA sequencing followed by prediction of potentially expressed proteins based on predicted genes. A major weakness of all database approaches is that they are unable to identify peptides that are not contained in the database. This is a particularly important limitation in the case of microbial peptides. It is well known that only 1%-10% of all microbes found in the environment can be cultured [16]–[18]. Thus there are many bacteria that have not been previously identified. Indeed, even among culturable organisms many remain uncharacterized due to extreme diversity. Even in the case where a microbial genome has been sequenced, it is quite possible that the strain under investigation exhibits amino acid mutations relative to the peptides in the database. Only de novo sequencing offers the possibility of identifying novel peptides.

Since de novo sequencing does not depend on a database of known peptides it offers the possibility of identifying novel peptides. It also offers the possibility of studying a proteome before the genome has been sequenced. The key problems involved in de novo sequencing are those of identifying the subset of peaks in the spectrum that specify an ion ladder and then determining the sequence of amino acids that are most consistent with the ion ladder. The b-ion ladder consists of those peaks that correspond to the prefixes of the peptide. In contrast, the y-ion ladder consists of those peaks that correspond to the suffixes of the peptide. If an ion ladder is complete then differences in adjacent peaks indicate the m/z value of the amino acid that distinguishes those adjacent peaks. It is this information that is used to determine the peptide sequence.

The approaches that most de novo algorithms take to evaluating candidate peptides mirror the two major approaches used by databases search methods. One approach is cross-correlation [5], [13], [19]. The cross-correlation function measures the coherence of the experimental spectrum with the virtual spectrum of a peptide candidate. The other common approach is based on probabilities. Probabilistic approaches assign probabilities to observed fragment peaks and then combine these for an overall peptide score [7], [9], [10], [20]. As in the case for database searches, the goal is to distinguish between significant peaks and noise peaks in the spectrum [21], [22]. In the case of PepNovo and PepNovo+ the probabilistic scoring function is in the form of a likelihood ratio hypothesis test [10], [23]. The Sherenga algorithm uses a scoring function that is based on a likelihood test that compares two possible explanations for the observed peaks [9]. In the first explanation, the peaks are evaluated as a result of the fragmentation process. This is based on a model that describes

J. R. Rose and J. P. Cleveland are with the Department of Computer Science and Engineering, University of South Carolina, Columbia South Carolina 29208. email: rose@cse.sc.edu, jimmycleveland@gmail.com

A. Fox is with the Department of Pathology, Microbiology and Immunology, University of South Carolina School of Medicine, Columbia South Carolina 29209. email: alvin.fox@uscmed.sc.edu

the probabilities of detecting certain fragment types in the spectrum. In the second explanation, the peaks are modeled as having been created by a random process. Havilio et al. added to this the consideration of correlation between intensity levels of different fragments [20]. Additionally, they treated fragment type, fragment mass, and fragment mass/peptide mass as special cases of fragments' chemical properties [20]. Frank and Pevzner extended these ideas through the use of a probabilistic network to model fragmentation [10]. Their score is computed as a log-likelihood that measures how likely it is that there was a cleavage of the peptide at a given mass.

II. Consideration of Amino Acid Usage

One important piece of information that is missing from current probabilistic and cross-correlation scoring function is the prior distribution of amino acid usage. This distribution describes the percentage of each amino acid as well as the probability of combinations of amino acids in peptide sequences. It captures the mutual information present in adjacent residues in the protein sequences from which the distribution was derived. By leaving this information out, one is effectively using a flat prior that treats all combination of amino acids as equally likely. It is clearly not the case that amino usage in peptides is random. NovoHMM is an interesting exception. Although NovoHMM uses a hidden Markov model instead of likelihood model, it implicitly incorporates information concerning amino acid usage by training with spectrum/peptide pairs [12]. However, NovoHMM's understanding of amino acid usage is inherently limited since it is derived entirely from available spectrum/peptide training pairs.

Researchers have recognized that there is bias in the types of peptides that are consistently observed by current MS/MS technology. These preferentially observed peptides are called proteotyptic peptides [24]–[26]. In this case, the bias is not a reflection of the proteome signature but of the experimental protocol and MS/MS technology. PepNovo+ employs a ranking algorithm to rerank candidate peptides produced by its fragmentation model. While the PepNovo+ ranking algorithm considers sequence composition features, it is limited to amino acid triplets that are compiled from proteotypic sequences. The result is a single distribution describing the proteotypic character of triplets averaged over all such training sequences [23].

In contrast, the QuasiNovo scoring function described in this paper recognizes that amino acid usage can vary widely from organism to organism [27], [28]. Typically it is similar between closely related taxa but can be quite different when taxa are distantly related. Consequently, QuasiNovo's understanding of amino acid usage is provided by several models. These models are derived from protein sequence data alone. This data is much more plentiful and accurate than spectrum/peptide pairs and leads to a more detailed and nuanced understanding of amino acid usage. In this paper we present results supporting the hypothesis that a scoring function that takes amino acid usage into account can significantly improve the accuracy of peptides derived via de novo sequencing.

III. Methods and Data

Our investigations were designed to evaluate the utility of a scoring function based on amino acid usage distributions. These distributions were created by selecting a number of proteomes from which to compile a composite amino acid distribution. (In practice, we started with translations from genome sequences.) The proteins from the selected proteomes were processed in the following manner. First, we chose a tuple length L. We then tabulated the frequency of occurrence of each tuple using a sliding window of length L. Let $< a_1a_2 \ldots a_n >$ be a contiguous sequence of n amino acids. There are $n - L + 1$ tuples of length L in this sequence: $< a_1a_2 \ldots a_L >$, $< a_2a_3 \ldots a_{L+1} >$, ..., $< a_{n-L+1}a_{n-L+2} \ldots a_n >$. Our preliminary studies were conducted using a tuple length of 6.The justification for choosing this length is that we expected a longer tuple to be more robust in the case of missing peaks. However, the number of unique tuples is exponential in the length of the tuple, *i.e.*, 20^L (since there are 20 amino acids). Thus anything larger than 6 becomes unmanageable. Additionally, it is difficult to collect enough peptide data to adequately populate a larger table. Even in the case of tuples of length 6, not all combinations of length 6 are observed so it is necessary to initialize those entries to some small epsilon value. Finally, the frequencies are then normalized to give the probability of each tuple in the composite set of peptides. From these 6-tuples we can also derive conditional probabilities of the form $p(a_6|a_1a_2a_3a_4a_5)$.

The amino acid distribution models amino acid usage and can be used to estimate the probability of observing an amino acid sequence. This model is used to compute the probability of a peptide of n amino acids by taking the product of the probability of the first tuple of length $L - 1$ times the subsequent $n - L + 1$ overlapping conditional probabilities based on tuples of length L in the peptide, *i.e.*,

$$P_{AAU}(P|M_{AAU}) = p(P_{1,L-1}) \prod_{i=L}^{n} p(P_i|P_{i-L+1,i-1}) \quad (1)$$

In this equation $p(P_{1,L-1})$ is the probability of the first $L - 1$ residues in peptide P and $p(P_i|P_{i-L+1,i-1})$ is the conditional probability of the ith amino acid given the preceding $L - 1$ amino acids. The probabilities $p(P_{1,L-1})$ and $p(P_i|P_{i-L+1,i-1})$ are defined by the amino acid usage model M_{AAU}, *i.e.*, the normalized amino acid distribution.

If a de novo sequencing algorithm with this type of scoring function could be shown to be competitive with existing de novo sequencing algorithms then one would expect a model that combined a probabilistic fragmentation model with an amino acid usage prior to perform substantially better than one using an implicit flat prior. To this end, we selected the same data set of 280 spectra used by Frank and Pevzner [10]. They used this data set to compare PepNovo with Sherenga, PEAKS, and Lutefisk. This data set comes from two sources, the ISB protein mixture data set [29] and the Open Proteomics Database (OPD) [30]. In this data set, peptides average 10.5 residues in length. Frank and Pevzner demonstrated that PepNovo outperformed Sherenga, Peaks and Lutefisk on this data set. This data set was also used by Fischer et al. to compare

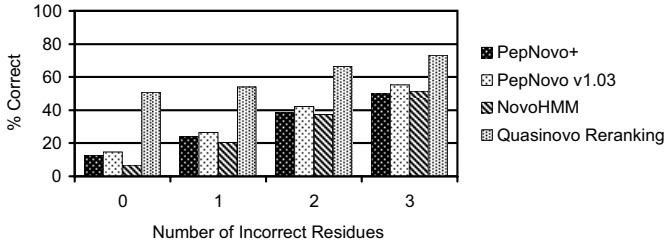

Fig. 1. Results for set of 280 MS/MS test spectra comparing PepNovo+, PepNovo, NovoHMM, with a QuasiNovo reranking.

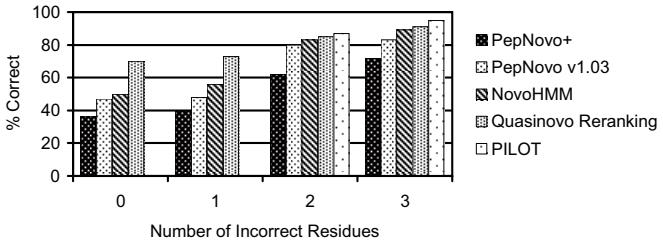

Fig. 3. Results for set of 100 MS/MS test spectra comparing PepNovo+, PepNovo, NovoHMM, PILOT and QuasiNovo, and experimental scoring function based on amino acid usage

NovoHMM with PepNovo, Sherenga, PEAKS, and Lutefisk [12]. In their study NovoHMM outperformed its competitors. Consequently, the focus of our evaluation was a comparison of the results of our scoring function versus PepNovo and NovoHMM.

The 280 spectra in the Frank-Pevzner data set are comprised of spectra from 174 *Escherichia coli* peptides, 27 *Mycobacterium smegmatis* peptides, 67 *Bos taurus* peptides, and 12 *Homo sapiens peptides*. The three major categories represented in this data set are *Gammaproteobacteria* (*E. coli*), *Actinobacteria* (*M. smegmatis*), and *Mammalia* (*B. taurus* and *H. sapiens*). Amino acid distributions were constructed for each of the 3 categories. *E. coli* and *M. smegmatis* peptides were specifically excluded from their respective distributions to demonstrate the ability of sequencing novel peptides. The *Gammaproteobacteria* distribution was constructed from approximately 23 million tryptic peptides from 205 gammaproteobacterial proteomes not including *E. coli*. The *Actinobacteria* distribution was constructed from approximately 7 million tryptic peptides from 57 complete actinobacterial genomes, not including *M. smegmatis*. Similarly, two mammalian distributions were created, one excluding *H. sapiens* and the other excluding *B. taurus*. The mammalian distribution used to score *H. sapiens* peptides was constructed from the complete proteomes of *B. taurus*, *R. norvegicus*, and *M. musculus*. The distribution used to score *B. taurus* peptides was constructed from complete proteomes of *H. sapiens*, *R. norvegicus*, and *M. musculus*.

IV. Experimental Results

Initial results are shown in Figure 1. The common practice in the de novo sequencing literature of presenting results in terms of the number of predictions that are correct within one, two, and three amino acids was followed. Each category is cumulative, *e.g.*, the category correct within 3 residues also includes the number of peptides with fewer errors. This figure depicts the accuracy of the top scoring candidate as selected by each method. Both PepNovo and NovoHMM produce a single top scoing candidate. In contrast, using default settings PepNovo+ produces 50 peptides sorted by rank. The QuasiNovo scoring function was used to rescore candidate peptides produced by PepNovo, PepNovo+, and NovoHMM. For each spectrum, a set comprised of the top 50 candidates produced by PepNovo+ and the single candidates produced by PepNovo and NovoHMM was created. The QuasiNovo scoring function was then used to select the peptide that produced the highest resulting score from this set. These results are labeled QuasiNovo Reranking in Figure 1. The most striking feature of the results presented in Figure 1 is that the QuasiNovo Reranking scores significantly higher than do PepNovo, PepNovo+ and NovoHMM. Recall that this reranking entails taking the 50 peptides suggested by PepNovo+ and the single peptides suggested by PepNovo and NovoHMM and then selecting the peptide with the highest QuasiNovo score. These results indicate that amino acid usage carries conditioning information about protein sequences that provides additional precision in mapping from the spectrum to the corresponding peptide.

A common alternative performance metric in the de novo sequencing literature is to present results in terms of the percentage of correct contiguous subsequences [10], [12], [13]. Not all de novo sequencing algorithms predict complete peptides. Often the peaks near the terminal ends are weak or missing. Consequently, the correct subsequences tend to be in the middle of the peptide. Figure 2 presents the longest subsequence results for the Frank-Pevzner dataset of 280 spectra. These results were derived by first finding the longest correct subsequence in the data set for each algorithm and then tallying the counts for each length. In this figure, the curve corresponding to the QuasiNovo reranking dominates the other curves by a significant amount over all subsequence lengths of four and greater.

DiMaggio and Floudas used 100 spectra from the Frank-Pevzner data set [10] to compare PILOT with PepNovo, and EigenMS. In their study [13], the PILOT results were slightly better than those of PepNovo and EigenMS. We evaluated a QuasiNovo reranking of the peptides proposed by PepNovo+ and NovoHMM for this set in order to see what effect the consideration of amino acid usage would have. The results of the reranking are shown in Figure 3. Notice that the results for PILOT only indicate the number of peptides (out of 100) that are correct within 2 amino acids and within 3 amino acids. This is because DiMaggio and Floudas did not publish results for completely correct peptides. It is instructive to note that the QuasiNovo reranking of the PepNovo and NovoHMM results increase the number of completely correct peptides from 47 and 50, respectively, to 72. Finally, even in the case of peptides considered correct within 3 amino acids where PILOT achieves 95 out of 100, the QuasiNovo reranking results are 93 out of 100, *i.e.*, comparable results.

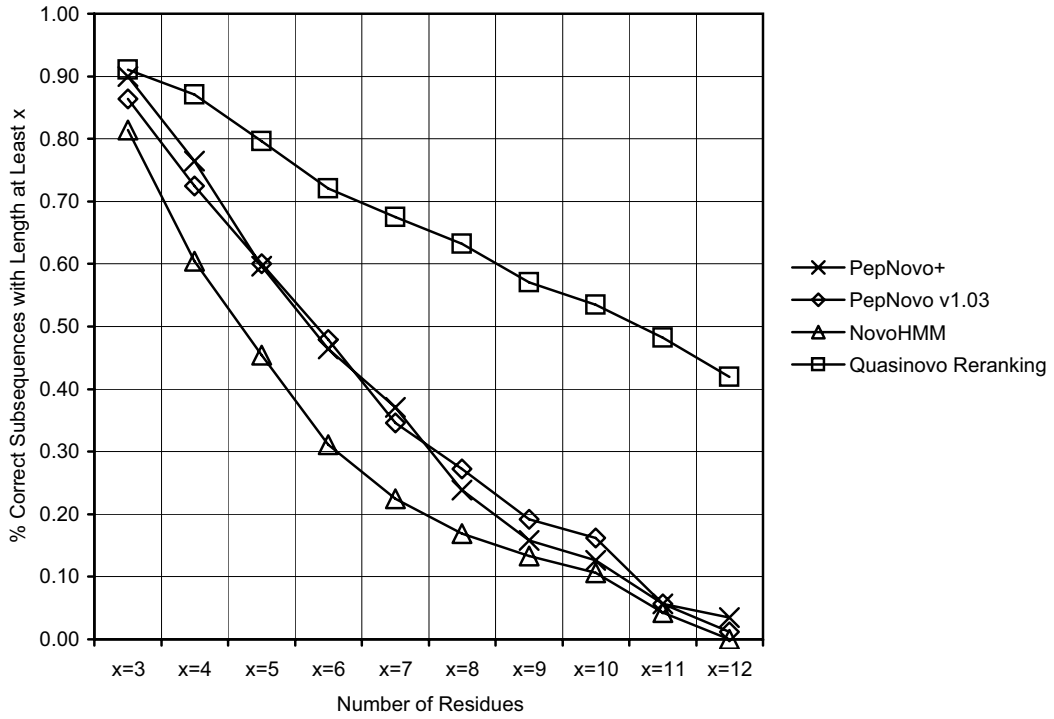

Fig. 2. Cumulative results for set of 280 MS/MS test spectra illustrating the proportions of predictions that had a correct subsequence of length at least x, for $3 \leq x \leq 12$.

The results in Figure 3 assume isobaric residues to be equivalent since DiMaggio and Floudas published these statistics but did not publish the actual peptides proposed by PILOT for this data set. Specifically, this means that the pairs I/L and Q/K are treated as identical amino acids. For example, the assignment of an isoleucine in the candidate peptide where the actual peptide contains a leucine is considered correct. This is a common practice since de novo sequencing algorithms that do not take amino acid usage into account have no basis for distinguishing between isobaric residues. Since QuasiNovo models amino acid usage, its scoring function is able to distinguish among isobaric residues. Consequently, QuasiNovo selects the residue with the highest probability in the context of a given peptide. Another weakness of methods that do not consider amino acid usage lies in how they treat missing peaks. This commonly occurs when the b1-ion (corresponding to the N-terminal amino acid) is missing from the spectrum. Peaks corresponding to other b- or y-ions may also be missing from the spectrum. Since peaks corresponding to b1-ions are frequently missing, more errors would be expected in the prediction of this terminal residue. If a peak corresponding to a b1-ion is missing from the spectrum then a de novo sequencing algorithm must make a prediction based on the next peak in the ion ladder, *i.e.*, the b2-ion. Table I shows the accuracy of the predictions made by PepNovo+, NovoHMM, and the QuasiNovo reranking for terminal ion pairs in the Frank-Pevzner dataset of 280 spectra. Table I does not assume isobaric equivalence. The values in the table were derived by tallying the number of correctly predicted terminal pair residues. For example, the values in the b2-ion column were

TABLE I
COMPARISON OF TERMINAL PAIR AND OVERALL ACCURACY

	terminal ion pair		complete peptide
algorithm	b_2-ion	y_2-ion	
PepNovo+	0.509	0.616	0.702
NovoHMM	0.523	0.759	0.735
QuasiNovo Reranking	0.716	0.813	0.815

determined by summing the number of correctly predicted residues in the first two positions in the 280 peptides and then dividing by 560, *i.e.*, 2 residues * 280 peptides. As shown in Table I, the QuasiNovo reranking results are superior to those of PepNovo+ and NovoHMM for predicting the correct residue pairs corresponding b2-ions and y2-ions. Notice that the accuracy of the amino acids predicted for the y2-ion for all algorithms in Table I are closer to the accuracy for the complete peptide than they are to the b2-ion. The y1-ion is not as frequently missing from the spectrum as the b1-ion.

When a peak is missing and can not be inferred, methods that do not model amino acid usage are typically able to propose a combination of residues for that part of the peptide. However, they are not able to specify the particular order in which the combination of residues appear in the peptide. It is for this reason that it has become common practice to present results in terms of percentage of predictions that are correct within one, two, and three amino acids as shown in Figures 1 and 3. In contrast, QuasiNovo uses its model of amino acid usage to distinguish between possible permutations. On this basis it selects the permutation with the greatest probability.

The results in Table I as well as the preceding figures

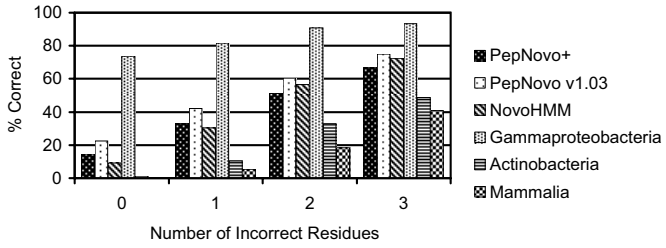

Fig. 4. Results for set of 76 MS/MS test spectra for E. coli peptides comparing PepNovo+, PepNovo, NovoHMM, with three QuasiNovo scoring functions based on amino acid distributions in *Gammaproteobacteria*, *Actinobacteria*, and *Mammalia*.

demonstrate the utility of integrating amino acid usage considerations in a scoring function. An obvious question is what influence the choice of proteomes used to build the proteome signatures has on the accuracy of the peptide scoring function. The 280 spectra in the Frank-Pevzner data set are comprised of 3 major categories: *Gammaproteobacteria*, *Actinobacteria*, and *Mammalia*. The experiment shown in Figure 3 was repeated. However, this time three different proteome-signature-based scoring functions were used to evaluate the subset of 76 *E. coli* peptides from the original 100 peptides. The scoring function labeled *Gammaproteobacteria* was compiled using only amino acid usage data from *Gammaproteobacteria* proteomes. Similarly, the scoring functions labeled *Actinobacteria* and *Mammalia* were derived exclusively from amino acid usage data from *Actinobacteria* and *Mammalia*, respectively. Given that all 76 peptides are from *E. coli*, it is no surprise that the results for the scoring function derived from gammaproteobacterial peptides are significantly higher than the other two scoring functions as shown in Figure 4. This data hints at the sensitivity of the accuracy of the QuasiNovo scoring function to the peptide data from which it is constructed. It should also be noted that the results shown in Figure 4 do not assume isobaric residues to be equivalent. One of the strengths of considering amino acid usage is that it provides a statistical basis for choosing between isobaric equivalent residues.

These results show that amino acid usage can be used as prior information to improve significantly the accuracy of the scoring functions used by current de novo sequencing algorithms. They also support the hypothesis that a significant additional increase in sequencing accuracy could be attained by including consideration of amino acid usage as an integral component of a scoring function.

V. Conclusion

The results of our investigations conclusively demonstrate two results. First, when we use a proteome signature-based scoring function to re-rank a combination of PepNovo+'s, PepNovo's, and NovoHMM's candidate peptides, the peptide that gets our highest score demonstrates significant improvement in accuracy as compared to PepNovo+'s, PepNovo's, and NovoHMM's first choice. Second, top-performing de novo sequencing programs such as PepNovo+ are able to generate good quality candidate peptides. In addition, they are able to compute candidate peptides very efficiently. For example, DiMaggio and Floudas report that PILOT takes 5-20 seconds to evaluate a spectrum on an Intel Pentium 4 3.0GHz Linux-based computer [13]. Even so, they are often not able to correctly rank them. As a consequence, a suboptimal candidate is selected as the highest ranking peptide and the accuracy is considerably lower than it should be. Put simply, scoring functions that do not consider amino acid usage appropriately are often not able to select the most correct peptide from a pool of candidates. Our results demonstrate the utility of combining fragmentation and proteome signature scoring functions.

We are currently developing scoring functions that integrate proteome signature models so that amino acid usage is considered during the process of candidate peptide generation. In the proof-of-concept approach presented in this paper, amino acid usage was considered after candidate peptides had been generated. It is possible that better candidates may have been discarded or not considered at all during the candidate generation phase. We expect consideration of amino acid content during candidate generation to further improve de novo peptide prediction accuracy.

The comparison of amino acid usage models in Figure 4 shows that the choice of amino acid usage model is important. Consequently, this is another important area of investigation. The results of the *Gammaproteobacteria* model in Figure 4 are impressive when compared with those of PepNovo, PepNovo+ and NovoHMM. The *Gammaproteobacteria* amino acid usage model in Figure 4 was compiled by aggregating data from the 205 proteomes from the *Gammaproteobacteria* class. Even the models constructed for mammalian peptides were not particularly focused, containing data from *H. sapiens*, *B. taurus*, *R. norvegicus*, and *M. musculus*. It is reasonable to expect that the accuracy of the scoring function will be improved by creating statistical models of amino acid usage that are closer to the proteome signature of the peptides under consideration. It is hypothesized that more focused models at the level of family, and genus will demonstrate greater improvements in accuracy relative to the results presented here.

One argument for pursuing de novo sequencing is the ability to sequence peptides expressed from unsequenced genomes. In the case of such a peptide, it is not possible to have the actual amino acid usage model. However in the case of bacteria, simple physiological tests (*e.g.* Gram stain) for cultured organisms can help limit the data set or limit the taxonomic categories under examination. For un-cultured single cells, equivalent information may also be obtainable. In both cases, it is possible to use universal primers to extract small subunit ribosomal RNA and sequence rRNA genes without having to sequence the entire genome. SSU rRNA databases are already the main source of microbial diversity information owing to rRNAs' role as the gold standard for microbial identification [31]. While the high degree of conservation of rRNA genes reduces their usefulness in resolving fine details at the strain or species level, it nevertheless makes them useful for inference of deeper phylogeny [32]. Several published studies show that 16S rRNA genes provide genus identification in over 90% of the cases considered [33], [34]. This information can be used to select the most appropriate available amino acid

usage model, including flat priors if only distantly related taxa are known. Consequently, it is important to compile and investigate models at the taxonomic levels of genus and family. It is currently unknown how much improvement more focused models will produce. If the improvement is limited then more general models might be preferred. We are currently investigating the extent to which more focused amino acid usage models improve the accuracy of the scoring functions that use such data.

ACKNOWLEDGMENTS

This research was supported by a grant from the Sloan Foundation Indoor Air Program. The research was also suported by award #0959427 from the National Science Foundation The experiments were run on an SGI Altix 4700 system, with 128 computing cores and 256GB shared memory funded by NSF award #0708391.

REFERENCES

[1] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, pp. 198–207, 2003.

[2] R. D. Smith, G. A. Anderson, M. S. Lipton, L. Pasa-Tolic, Y. Shen, T. P. Conrads, T. D. Veenstra, and H. R. Udseth, "An accurate mass tag strategy for quantitative and high-throughput proteome measurements," *Proteomics*, vol. 2, pp. 513–523, 2002.

[3] D. A. Wolters, M. P. Washburn, and J. R. I. Yates, "An automated multidimensional protein identification technology for shotgun proteomics," *Anal. Chem.*, vol. 73, pp. 5683–5690, 2001.

[4] D. N. Perkins, D. J. C. Pappin, D. M. Creasy, and J. S. Cottrell, "Probability-based protein identification by searching sequence databases using mass spectrometry data," *Electrophoresis*, vol. 20, pp. 3551–3567, 1999.

[5] J. K. Eng, A. L. McCormack, and J. R. I. Yates, "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database," *J. Am. Soc. Mass Spectrom.*, vol. 5, pp. 976–989, 1994.

[6] J. I. Yates, J. K. Eng, A. L. McCormack, and D. Schieltz, "A method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database," *Anal. Chem.*, vol. 67, pp. 1426–1436, 1995.

[7] V. Bafna and N. Edwards, "Scope: a probabilistic model for scoring tandem mass spectra against a peptide database," *Bioinformatics*, vol. 17, pp. S13–S21, 2001.

[8] J. A. Taylor and R. S. Johnson, "Sequence database searches via de novo peptide sequencing by tandem mass spectrometry," *Rapid Commun. Mass Spectrom.*, vol. 11, pp. 1067–1075, 1997.

[9] V. Dancik, T. A. Addona, K. R. Clauser, J. E. Vath, and P. A. Pevzner, "De novo peptide sequencing via tandem mass spectrometry," *J Comp Biol.*, vol. 6, pp. 327–342, 1999.

[10] A. Frank and P. Pevzner, "Pepnovo: De novo peptide sequencing via probabilistic network modeling," *Anal. Chem.*, vol. 77, pp. 964–973, 2005.

[11] M. Bern and D. Goldberg, "De novo analysis of peptide tandem mass spectra by spectral graph partitioning," *J Comp Biol.*, vol. 13, pp. 364–378, 2006.

[12] B. Fischer, V. Roth, F. Roos, J. Grossmann, S. Baginsky, P. Widmayer, W. Gruissem, and J. M. Buhmann, "Novohmm: A hidden markov model for de novo peptide sequencing," *Anal. Chem.*, vol. 77, pp. 7265–7273, 2005.

[13] P. A. DiMaggio and C. A. Floudas, "De novo peptide identification via tandem mass spectrometry and integer linear optimization," *Anal. Chem.*, vol. 79, pp. 1433–1446, 2007.

[14] K. R. Clauser, P. Baker, and A. L. Burlingame, "Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing ms or ms/ms and database searching," *Anal. Chem.*, vol. 71, pp. 2871–2882, 1999.

[15] M. Mann and M. Wilm, "Error-tolerant identification of peptides in sequence databases by peptide sequence tags," *Anal. Chem.*, vol. 66, pp. 4390–4399, 1994.

[16] D. M. Ward, W. R., and M. M. Bateson, "16s rrna sequences reveal numerous uncultured microorganisms in a natural community," *Nature*, vol. 345, pp. 63–65, 1990.

[17] U. B. Goebel, "Phylogenetic amplification for the detection of uncultured bacteria and the analysis of complex microbiota," *J. Microbiol. Methods*, vol. 23, pp. 117–128, 1995.

[18] J. M. Gonzalez and C. Saiz-Jimenez, "Application of molecular nucleic acidbased techniques for the study of microbial communities in monuments," *Int. Microbiol.*, vol. 8, pp. 189–194, 2005.

[19] Y. Fu, Q. Yang, R. Sun, D. Li, R. Zeng, C. X. Ling, and W. Gao, "Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry," *Bioinformatics*, vol. 20, pp. 1948–1954, 2004.

[20] M. Havilio, Y. Haddad, and Z. Smilansky, "Intensity-based statistical scorer for tandem mass spectrometry," *Anal. Chem.*, vol. 75, pp. 435–444, 2003.

[21] R. G. Sadygov and J. R. Yates, "A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases," *Anal. Chem.*, vol. 75, pp. 3792–3798, 2003.

[22] T. Fridman, J. Razumovskaya, N. Verberkmoes, G. Hurst, V. Protopopescu, and Y. Xu, "The probability distribution for a random match between an experimental-theoretical spectral pair in tandem mass spectrometry," *J. Bioinform. Comput. Biol.*, vol. 3, pp. 455–476, 2005.

[23] A. M. Frank, "A ranking-based scoring function for peptide-spectrum matches," *J. Proteome Res.*, vol. 8, pp. 2241–2252, 2008.

[24] R. Craig, J. Cortens, and R. Beavis, "The use of proteotypic peptide libraries for protein identification," *Rapid Commun. Mass Spectrom.*, vol. 19, pp. 1844–1850, 2005.

[25] H. Tang, R. Arnold, P. Alves, Z. Xun, D. Clemmer, M. Novotny, J. Reilly, and P. Radivojac, "A computational approach toward labelfree protein quantification using predicted peptide detectability," *Bioinformatics*, vol. 22, pp. e481–e488, 2006.

[26] J. Ranish, B. Raught, R. Schmitt, T. Werner, K. B., and R. Aebersold, "Computational prediction of proteotypic peptides for quantitative proteomics," *Nat. Biotechnol.*, vol. 25, pp. 125–131, 2007.

[27] P. Foster and D. A. Hickey, "Compositional bias may affect both dna-based and protein-based phylogenetic reconstructions," *J. Mol. Evol.*, vol. 48, pp. 284–290, 1999.

[28] G. A. C. Singer and D. A. Hickey, "Nucleotide bias causes a genomewide bias in the amino acid composition of proteins," *Mol. Biol. Evol.*, vol. 17, pp. 1581–1588, 2000.

[29] A. Keller, S. Purvine, A. I. Nesvizhskii, S. Stolyar, D. R. Goodlett, and E. Kolker, "Experimental protein mixture for validating tandem mass spectrometry analysis," *OMICS J. Integr. Biol.*, vol. 6, pp. 207–212, 2002.

[30] J. T. Prince, M. W. Carlson, R. Wang, P. Lu, and E. M. Marcotte, "The need for a public proteomics repository," *Nat. Biotechnol.*, vol. 22, pp. 471–472, 2004.

[31] N. Pace, "Mapping the tree of life: Progress and prospects," *Microbiology and Molecular Biology Reviews*, vol. 73, no. 4, pp. 565–576, December 2009.

[32] J. M. Janda and S. L. Abbott, "16s rrna gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls," *Journal of Clinical Microbiology*, vol. 45, no. 6, pp. 2761–2764, September 2007.

[33] M. Drancourt, C. Bollet, A. Carlioz, R. Martelin, J. Gayral, and D. Raoult, "16s ribosomal dna sequence analysis of a large collection of environmental and clinical unidentifiable bacterial isolates," *Journal of Clinical Microbiology*, vol. 38, no. 10, pp. 3623–3630, October 2000.

[34] S. Mignard and J. P. Flandrois, "16s rrna sequencing in routine bacterial identification: a 30-month experiment," *Journal of Microbiological Methods*, vol. 67, no. 3, pp. 574–581, December 2006.