A MACHINE LEARNING APPROACH TO PREDICTING PEPTIDE FRAGMENTATION SPECTRA

RANDY J. ARNOLD,[1] NARMADA JAYASANKAR,[2] DIVYA AGGARWAL,[2]
HAIXU TANG,[2,3] AND PREDRAG RADIVOJAC[2*]

1) Department of Chemistry, Indiana University, Bloomington, IN 47405
2) School of Informatics, Indiana University, Bloomington, IN 47408
3) Center for Genomics and Bioinformatics, Indiana University, Bloomington, IN 47405
**Corresponding author*

Accurate peptide identification from tandem mass spectrometry experiments is the cornerstone of proteomics. Although various approaches for matching database sequences with experimental spectra have been developed to date (e.g. Sequest, Mascot) the sensitivity and specificity of peptide identification have not yet reached their full potential. This is in part due to the tradeoffs between robustness and accuracy of the existing methods with respect to the non-uniform nature of peptide fragmentation and bond cleavages induced by different mass spectrometers. Accordingly, it is expected that new approaches to *de novo* predicting peptide fragmentation spectra will enable more accurate peptide identification. To address this problem, here we used a data-driven approach to learn peptide fragmentation rules in mass spectrometry, in the form of posterior probabilities, for various fragment-ion types of doubly and triply charged precursor ions. We show that the accuracy of our neural-network based methodology is useful for subsequent peptide database searches and that the most useful rules of fragmentation significantly differ across ion and precursor types.

1 Introduction

Recent advances in separations and mass spectrometry have enabled a surge in the comprehensive analysis of cellular proteins, commonly referred to as proteomics.[1,2] The critical development in this area is the ability to identify a peptide, or in some cases entire proteins, from the fragment ions generated by tandem mass spectrometry.[3] Various dissociation methods have been introduced, including commonly used gas phase collision-induced dissociation (CID),[4] surface-induced dissociation,[5] photodissociation,[6,7] electron-capture dissociation,[8] and electron transfer dissociation.[9] The resulting tandem mass spectra are compared with *in silico*, i.e. computer generated, spectra derived from peptides in the available protein database.[10] The commonly used protein identification tools often use ad-hoc rules[11] or unified probabilistic models[12,13] to estimate the likelihood that a given experimental spectrum was generated from each sequence contained in the database. The final assignments and confidence levels are then based on both the scores and database content.[14,15]

In practice, the peptide fragmentation into various ions may differ for several reasons. For many peptides, cleavages of amide bonds dominate the fragmentation and produce a series of b- and y-ions. For other peptides, the enhanced fragmentation at some types of amino acid residues may dominate. Also, the charge carried by the precursor ion affects electron distribution along the cleaved peptide backbone.

Finally, fragmentation method and the energy level used in the experiments are also known to largely change the global behavior of peptide fragmentation.

The development of chemical theory of peptide fragmentation, e.g. the "mobile proton" model,[16-18] enabled the *de novo* prediction of fragmentation spectra from peptide sequences. Using a kinetic model, Zhang made the first successful attempt at predicting the low-energy CID spectra of singly and doubly charged peptides.[19] He recently introduced a simplified model that can accommodate peptides with three or more charges as well as sequences of increased lengths.[20] However, it is not clear how this approach could be extended to the other types of mass spectrometry instruments.

An ability to obtain large amounts of peptide fragmentation data relatively cheaply sprung the development of data-driven approaches and machine learning techniques. Elias et al.[21] were first to successfully utilize a set of well annotated fragmentation spectra acquired from an electrospray ion-trap mass spectrometer in an attempt to infer the probabilistic rules of fragmentation. As a proof of concept, they learned a decision tree for the b- and y-ion fragmentation of the doubly charged precursors and used their model to significantly improve on Sequest scores of tryptic peptides. In addition, Elias et al. confirmed previously known rules of peptide fragmentation and presented a large set of new ones.

In this paper, we extend this approach to the triply charged precursors in addition to other, harder-to-predict, ion types (b–H_2O, b–NH_3, b–H_2O–NH_3, etc.). We note that it is not a trivial extension for two reasons. First, from the standpoint of protein identification, triply charged (+3) peptides seem to be much more difficult to identify than doubly charged (+2) peptides. For example, in the dataset used herein (Section 2.1), there were roughly four times as many doubly charged peptides (16,056) as triply charged peptides (4,130) that could be reliably identified by Mascot. This indicates that the current peptide identification tools may be better suited to the +2 charged peptides even though there may be a general preference to form +2, rather than +3 precursor ions. The distributions of fragment ions observed in +2 vs. +3 precursor ions are different. For instance, many +2 b- and y- fragment ions can be observed in the fragmentation spectra of the +3 precursors. For the +2 precursors, +1 b- and y-ions dominate the fragmentation spectra, whereas few, if any, of the +2 b- and y-ions can be observed. We illustrate these differences between the +2 and +3 precursors in Figure 1, where the same peptide was used to produce peptide fragmentation spectra. Second, the fragmentation mechanism of +3 ions is less understood than that of the doubly charged ions.[22] As a result, the new rules of fragmentation could be important since it is not as easy to develop a *de novo* prediction method for the +3 precursors as for the +2 precursors.[19]

The results of our study indicate that, for most of the ions, it is possible to predict the peptide fragmentation spectra with a useful accuracy. Furthermore, the obtained predictions can be used in a straightforward way to improve a simple correlation-based scoring function for peptide identification.

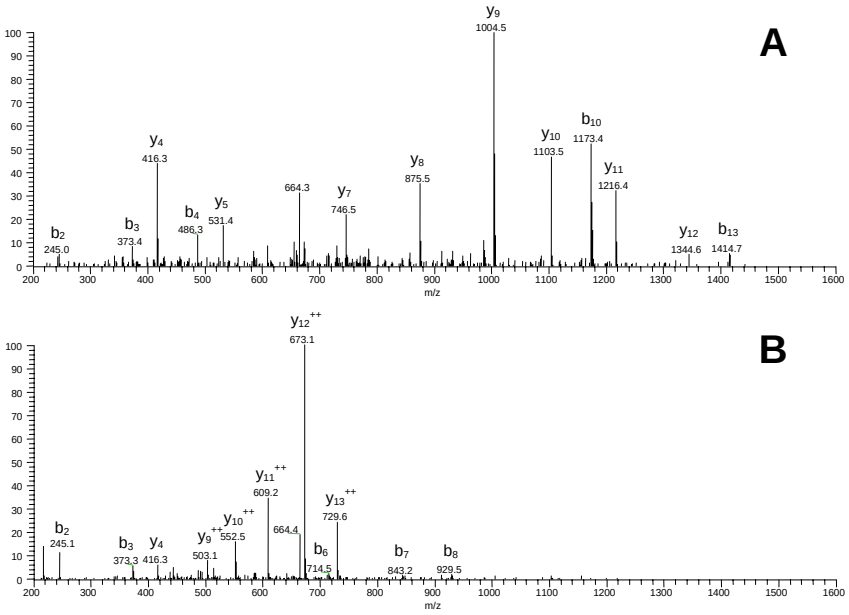

Figure 1. MS/MS spectra of the A) +2 and B) +3 precursor ions of peptide MLQLVEESKDAGIR acquired in consecutive scans of an LC-MS/MS experiment using an ion trap mass analyzer. Selected precursor *m*/*z* values are A) 795.2 and B) 530.69.

2 Methods

As previously mentioned, the major objective of this study was to use automated techniques in order to learn peptide fragmentation rules in the form of posterior probabilities and then utilize the trained model for peptide identification. The original problem of predicting spectral peak intensities was converted into a simpler and easy-to-interpret classification problem, in which the peak intensity was first normalized and then binned into two groups based on a threshold. Formally, given a precursor sequence S and its charge $q_S \in \{+2, +3\}$, we aimed to estimate the following set of probabilities: $P(I(i) \geq t \mid S, q_S)$, where $I(i)$ is the peak intensity of any fragment ion $i \in \{$*precursor*–H_2O, *b*–H_2O, *b*–NH_3, *b*–H_2O–NH_3, *y*–H_2O, *y*–NH_3, *y*–H_2O–NH_3, b^2, $y^2\}$ and t is an appropriately chosen threshold. In this study, by default, t is equal to 1% of the total intensity of the spectrum.

2.1 Datasets

Two groups of samples were prepared from isolated rat brains. The first group was produced by homogenizing hippocampus tissues and separating the lysate into four different fractions by differential sedimentation. The second group was produced by

separately lysing tissue from six different brain regions (amygdala, caudate putamen, frontal cortex, hippocampus, hypothalamus, and nucleus accumbens). All samples were digested separately with proteomics grade (modified) trypsin in the presence of an acid-labile surfactant. Tryptic peptides were separated by nano-flow reversed-phase liquid chromatography and electrosprayed directly into a ThermoFinnigan (San Jose, CA) LCQ Deca XP ion-trap mass spectrometer which recorded mass spectra and data-dependent tandem mass spectra of the peptide ions. By using dynamic exclusion, the mass spectrometer was limited to acquiring only one tandem mass spectrum for a given parent *m*/*z* over a 60-second window. Tandem mass spectra were filtered based on a total spectrum signal of 1 million counts for the first group of samples and 300 million counts for the second group of samples. All spectra were searched against protein sequences for *R. norvegicus* in the Swiss-Prot database[23] using Mascot[12] for peptide identification. Searches were performed with variable modifications of protein N-terminal acetylation and methionine oxidation selected and a maximum of one missed cleavage site. Mascot result files were parsed using a Protein Results Parser program written in-house to create a single training set with all peptides having Mascot scores of 40 or higher.

We normalized each spectrum to sum to one and divided all precursor peptides into doubly and triply charged. Peak intensities were estimated for the following ion types: *precursor*–H_2O, *b*, *b*–H_2O, *b*-NH_3, *b*–H_2O–NH_3, *y*, *y*–H_2O, *y*-NH_3, *y*–H_2O–NH_3, for the doubly charged precursors, while b^2 and y^2 ions were also considered for the triply charged ions. The set of precursor peptides of a given charge was filtered to prevent multiple copies. If two or more precursor peptides were identical, the one with the highest Mascot score over all fragment ions was retained. The dataset contained a significant number of identical precursor sequences and was reduced by factors of 9.3 and 6.3 for the doubly and triply charged peptides, respectively. The total counts of fragment ions corresponding to the set of unique precursor sequences are shown in Table 1.

A preliminary peptide identification using the new scoring based on the predicted fragment spectra was performed using the whole set of available proteins from *R. norvegicus*, containing 35,085 proteins (from the NCBI web site).

2.2 Data representation

To enable learning, each sequence fragment S was encoded into a fixed-length vector representation. More specifically, sequence $S = s_1s_2,\ldots s_n$ was represented by a vector of binary and real-valued features. Assuming the cleavage occurred between positions k and $k + 1$ in $S = s_1s_2,\ldots s_n$, the following features were constructed for all b- and y-ions: (i) amino acid compositions of the prefix subsequence $s_1s_2,\ldots s_k$ and the suffix subsequence $s_{k+1}s_{k+2},\ldots s_n$; (ii) lengths of both fragments, k and $n - k$; (iii) first neighbor prefix/suffix amino acids, s_k and s_{k+1}, and second neighbor prefix/suffix amino acids s_{k-1} and s_{k+2}; (iv) N- and C-terminal residues, s_1 and s_n; (v)

parent mass m; (vi) ion masses, m_{prefix} and m_{suffix}; and (vii) N-terminal acetylation. We also incorporated a number of features introduced by Elias et al.: gas phase basicity, helicity, hydrophobicity, and isoelectric point, both average and for the residues s_k and s_{k+1}.[21] Individual amino acids were encoded using a binary data representation,[24] expanded by adding oxidized methionine residue, while the compositional attributes were real-valued. To encode *precursor*–H_2O ions, we ignored the features related to the cleavage site. Overall, b- and y-ions were represented by 202 features, while the *precursor*–H_2O ion encoding contained 76.

Table 1. The total count of ions corresponding to the unique precursor sequences. An ion was considered present (positive cases) when its peak intensity exceeded 1% of the total spectral intensity. Otherwise, the ion was considered absent (negative cases).

Ion	*Doubly charged precursors*			*Triply charged precursors*		
	Positives	*Negatives*	*Total*	*Positives*	*Negatives*	*Total*
precursor – H_2O	239	1484	1723	64	590	654
b	5210	16916	22126	950	12000	12950
b – H_2O	1700	20426	22126	206	12744	12950
b – NH_3	678	21448	22126	117	12833	12950
b – H_2O – NH_3	249	21877	22126	121	12829	12950
b^2	-	-	-	1343	11607	12950
y	9323	12802	22126	1639	11311	12950
y – H_2O	431	21695	22126	132	12818	12950
y – NH_3	286	21840	22126	101	12849	12950
y – H_2O – NH_3	145	21981	22126	107	12843	12950
y^2	-	-	-	1953	10997	12950

2.3 Model selection and training

Predictors of ion intensities were built as ensembles of two-layer feed-forward neural networks, which, if provided with enough data, are known to be universal approximators of bounded functions.[25] A particularly useful property of these models is that the expected number of data points necessary for successful training is linear with the number of weights and that the training is relatively fast. Each model in an ensemble contained $h \in \{1, 2, 4, 8, 16, 32\}$ hidden neurons and one output neuron, all with sigmoidal activation function, and was trained using the resilient propagation algorithm.[26]

Since the threshold t for the quantization of the peak intensities was set to 1% of the total peak intensity, each resulting dataset was high-dimensional and class-imbalanced (even at 1% cutoff there were much fewer positives than negatives). Thus, we randomly under-sampled the majority class to the size of the positive class to train each network from a class-balanced dataset. However, to effectively use whole dataset, a different selection of negatives was made for each network in the ensemble.

Prior to network training, a t-test feature selection filter was employed to filter out unpromising features. The threshold for feature retention, t_f, was varied from the following set of values: {0.001, 0.01, 0.1, 1}. Clearly, in the case of $t_f = 1$ all features were retained. Finally, to remove correlated features we applied the principal component analysis and retained 95% of the variance. Feature selection thresholds and the number of hidden neurons were selected using a separate validation set for each individual model (20% of the training set), therefore producing only the final set of estimated accuracies. Each ensemble contained 30 neural networks.

2.4 Performance evaluation of the fragmentation ion peak prediction

A model was trained for each ion type separately using cross-validation. The non-overlapping folds were chosen at the level of precursor sequences since one +2 precursor ion of length n can produce a combination of the b-ions (b_1, b_2, etc.), y-ions (y_1, y_2, etc.) and their variants with neutral losses which could create information leak if distributed independently over training, validation, and test sets. Triply charged ions could also produce a combination of +2 b-ions (b_1^2, b_2^2, etc.) or +2 y-ions (y_1^2, y_2^2, etc.). Finally, the number of positive examples in each fold was balanced in order to achieve stable and realistic estimate of classification accuracy.

We measured sensitivity (*sn*) and specificity (*sp*) for each classifier. Sensitivity is defined as the percentage of positive examples, i.e. peaks over 1% of total intensity, correctly predicted, while specificity is the percentage of negative examples correctly predicted. We also estimated a class-balanced accuracy $acc = (sn + sp)/2$ and the area under the ROC curve (*AUC*), both of which are essentially unaffected by the disparity in class sizes. The area under the curve was calculated using the trapezoid rule.

2.5 Peptide identification based on the predicted spectra

The predicted fragment spectrum was constructed by combining the outputs of individual predictors for each ion type. Since individual predictors are trained on the class-balanced datasets, it is necessary to adjust their outputs according to the observed prior probabilities of positives and negatives for each ion type.[27] Given the class-balanced training set, an adjusted output score o_{adj} of the predictor can be calculated as

$$o_{adj} = \frac{p \cdot o}{1 - p + o \cdot (2p - 1)},$$

where $o \in [0, 1]$ is the output of a class-balanced trained model and p is the class prior for the positive examples. Different class priors were used for each fragment ion type.

We use the simple correlation coefficient to score the matching between an experimental spectrum $\{m_i^e, I_i^e\}$ and a predicted spectrum $\{m_j^p, I_j^p\}$, defined as

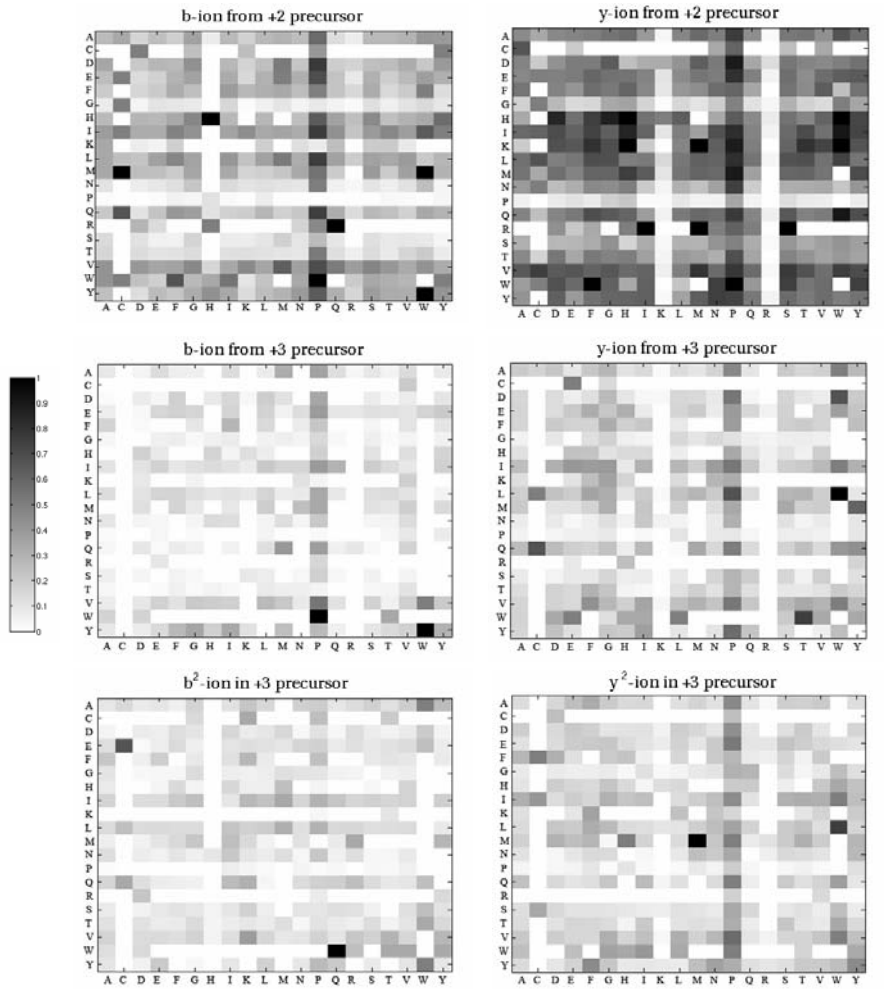

Figure 2. The amino acid preferences for peptide fragmentation. The frequencies of observing ion types (b-, y- or b^2, y^2) were plotted in grey scaling from 0 (white) to 1 (black). The rows indicate amino acid on the left-hand side, while the columns indicate amino acids on the right-hand side of the cleavage site.

$$s = \frac{\sum_{m_i^e \approx m_j^p} I_i^e \cdot I_j^p}{\sqrt{\sum_i (I_i^e)^2 \cdot \sum_j (I_j^p)^2}},$$

where m and I represent the mass and the intensity of each ion, respectively, and $m_i \approx m_j$ means that the difference between these two mass values is smaller than the tolerance t of the mass spectrometer ($t = 0.5$). For the comparison purposes, we also predict spectra using ad-hoc rules for each peptide (Section 3.3).

3 Experiments and Results

3.1 Analysis of amino acid preferences at the cleavage sites

Figure 2 illustrates the amino acid preferences for peptide fragmentation as influenced by the amino acids on both sides of the cleavage site. Several expected trends are observed across all ion types, specifically the preference for proline on the C-terminal side and the preference against proline on the N-terminal side of the cleavage site. The plots for the +2 precursors compare favorably with that shown previously[28] for cleavage intensity ratios for +2 peptides with partially mobile protons, even though the data shown here represent all proton mobility types (mobile, partially mobile, and non-mobile proton). Other interesting trends can be observed, such as the apparent preference for tryptophan (W) in the C-terminal position for +2 and +3 precursors, preference against glycine (G) and serine (S) in the N-terminal position for +2 precursors, and preference for cleavage between amino acid pairs LW (y-ions), WP and YW (b-ions), MM (y^2-ions), and WQ (b^2-ions) for +3 precursors. These trends are not fully understood at this time and may suggest previously unknown enhanced fragmentation sites. Note that Figure 2 does not account for the number of observed amino acid pairs at the cleavage site.

3.2 Evaluation of the peptide fragmentation prediction

Using the methodology presented in Sections 2.2-2.4, we trained classification models for twenty different ion types and evaluated performance of each model. In terms of varying learning parameters, we observed an increase in accuracy by 2-3 percentage points when ensembles of 30 models were used instead of a single model. In addition, the improvement of the non-linear models over the linear (a network with a single hidden neuron) was greater than 5 percentage points in some cases. We also note that in the case of b and y-ions, the number of selected hidden neurons reached its maximum (h = 32) indicating that an increase in dataset size and higher expressiveness of the classifier would likely cause an additional improvement in overall performance. The detailed evaluation of the classification accuracy appears in Table 2, while two sample predictions of triply charged precursors are shown in Figure 3.

The performance of our models was also evaluated against a decision tree model proposed by Elias et al.[21] We encoded the full set of features provided by the authors and trained a classifier for the *b* and *y* ions of the +2 precursors. We obtained classification accuracy of 73.5% (*sn* = 74.8%; *sp* = 72.1%) for the *b*-ion and 80.4% (*sn* = 82.3%; *sp* = 78.6%) for the *y*-ion, the two ion types studies by Elias et al. The differences between our ensemble models and decisions trees are statistically significant, with p-values below 0.01 in both cases (binomial distribution was used to calculate p-values). The C4.5 decision tree software was used with the default parameters.[29] More detailed evaluations can be found at our research home page.

Table 2. Classification accuracy [%] of the predictors on the doubly and triply charged ions; *sn* – true positive rate, *sp* – true negative rate, *acc* = (*sn* + *sp*)/2, *AUC* – area under the ROC curve.

Ion	*Doubly charged precursors*			*Triply charged precursors*		
	sn	*sp*	*acc/AUC*	*sn*	*sp*	*acc/AUC*
precursor – H_2O	72.0	60.8	66.4/70.7	81.3	68.5	74.9/79.7
b	80.4	75.4	77.9/85.8	80.6	71.9	76.3/84.6
b – H_2O	76.8	76.3	76.5/84.6	76.2	60.2	68.2/76.8
b – NH_3	75.8	76.0	75.9/82.8	76.9	65.0	70.9/78.6
b – H_2O – NH_3	69.1	64.6	66.8/73.1	81.8	51.9	66.9/68.1
b^2	-	-	-	88.4	75.8	82.1/88.5
y	84.7	79.3	82.0/89.5	88.9	79.1	84.0/91.4
y – H_2O	66.4	66.2	66.3/72.2	82.6	56.5	69.6/73.0
y – NH_3	70.3	70.8	70.6/79.0	81.2	59.8	70.5/77.8
y – H_2O – NH_3	60.7	51.1	55.9/56.5	83.2	54.3	68.7/69.6
y^2	-	-	-	87.9	72.6	80.2/86.8

3.3 Peptide identification experiments

The quality of peptide identification was estimated by comparing the correlation scores between the experimental spectrum and the computer generated spectra for the true precursor sequence and 500 spurious tryptic peptides selected from the rat proteome. We estimated (i) the average difference between the score for the true peptide and the score for the best scoring random peptide, and (ii) the average rank of the true peptide in the context of 500 candidates. The mass of the true peptide m was used to get candidate tryptic peptides in the rat proteome whose masses approximately matched the mass of the experimentally measured peptide (within $m-1$ and $m+3$). The candidate peptides were allowed to have up to one missed cleavage.

The evaluation of peptide identification is presented in Table 3. Our new scoring scheme was compared to an ad-hoc (also referred to as simple) scoring in which all possible *b* and *y* fragment ions are assigned intensities of 1, *b*–H_2O and *y*–H_2O are assigned peak intensities of 0.5, *b*–NH_3 and *y*–NH_3 are assigned peak intensities of 0.3, while *b*-H_2O–NH_3, *y*–H_2O–NH_3, b^2 and y^2 are assigned intensities of 0.1. Other fragment ions were assigned peak intensities of 0, as well as all fragment ions whose mass was greater than 2,000 due to the upper mass limit of the mass spectrometer. A negative score indicates that, on average, highest-scoring random peptides may be selected over the true peptides with higher confidence than the true peptides are selected over highest-scoring random peptides. These experiments provide evidence that our approach to spectrum prediction may provide highly promising peptide identification.

The spectra were classified into 8 categories according the chemical property of the peptides, namely +2 or +3 precursor, mobile (number of precursor charges is greater than total of H, K, and R) or non-mobile proton, and presence or absence of proline. While improvement over simple scoring is observed for all 8 categories, the

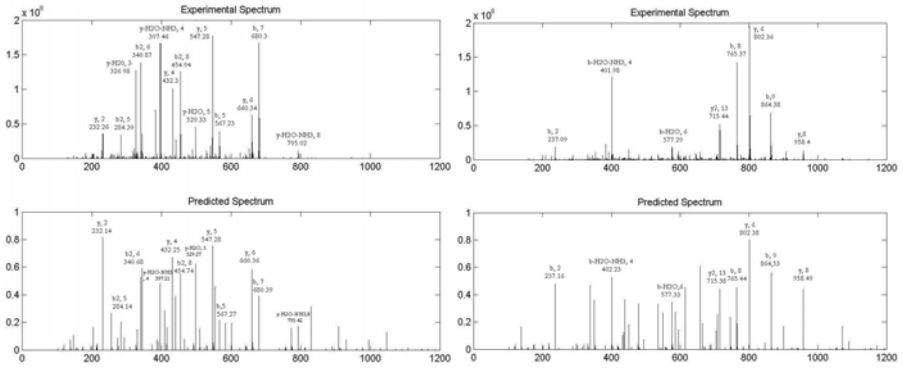

Figure 3. Experimental (upper panels) vs. predicted (lower panels) spectra for two triply charged precursor ions. The left panels correspond to sequence HRDTGILDSIGRZ, while the right panels correspond to the sequence HVLSGTLGCPEHTYR. Note that the first sequence corresponds to the non-mobile proton w/o proline case, while the second sequence corresponds to the non-mobile proton with proline case.

greatest improvement appears to be for triply-charged precursors, especially with a mobile proton and proline.

Table 3. The average difference in scores (*diff*) between the true peptide and the highest scoring random peptide and an average ranking of the true peptide (*rank*). To obtain scores and rankings we used 500 random peptides from the rat proteome having approximately the same precursor mass as the true peptide. The scores are separated for the cases of doubly vs. triply charged precursors, mobile vs. non-mobile proton, and presence vs. absence of proline (Pro) in the precursor sequence. Each field in the table was averaged using a set of 25 randomly taken precursor sequences (with 68% confidence intervals), identical for both scoring schemes.

Scoring scheme		*Doubly charged precursors*				*Triply charged precursors*			
		Mobile proton		*Non-mobile proton*		*Mobile proton*		*Non-mobile proton*	
		w/o Pro	*w Pro*	*w/o Pro*	*w Pro*	*w/o Pro*	*w Pro*	*w/o Pro*	*w Pro*
diff	*New*	.32±.03	.26±.04	.30±.03	.24±.04	.13±.03	.14±.04	.22±.03	.25±.04
	Simple	.22±.02	.14±.03	.23±.02	.15±.02	–.01±.02	–.03±.02	.08±.02	.09±.03
rank	*New*	1.1±0.1	1.4±0.2	1.1±0.1	1.5±0.2	1.8±0.7	1.5±0.2	1.4±0.4	1.2±0.2
	Simple	1.1±0.1	1.4±0.2	1.0±0.1	1.3±0.2	9.0±1.8	19.0±4.5	2.3±1.0	6.1±2.2

4 Discussion

Machine learning approaches have been extensively applied to proteomics research. They, however, mostly focused on either the preprocessing of the spectrum or the post-processing the peptide identification results of conventional tools. For examples, binary classifiers[30] and artificial neural networks[31] were introduced to evaluate the quality of MS/MS spectra before they were used for peptide identification; support vector machines were also used to classify the positive protein identification based on Sequest output.[32] Very little work, however, has addressed the potential of

applying machine learning to the peptide identification itself. We extend the previous work by Elias et al.[21] to the more challenging problem of predicting the full fragment spectra of peptides for both doubly- and triply-charged precursors. Our preliminary tests on the scoring of peptide identification showed encouraging results towards a new scoring scheme for peptide identification with better performance.

All classifiers used in this study were trained on balanced samples. Certainly, balanced training provides good insight into the class separability since it is not related to the relative fraction between the positive and negative datasets. On the other hand, application of such a predictor may cause significant overprediction if applied to the representative imbalanced dataset. This problem, however, was easily resolved by adjusting the outputs of the predictor, depending on the class priors. Our choice of neural networks compared to other machine learning techniques (SVMs or decision trees) were based on our experience with using the models and the fact that only minor modifications will be required in order to learn peak intensities in a regression-based approach.

The empirical rules we derived from machine learning approaches in this paper will be also useful for understanding the fragmentation mechanism of triply charged ions. The enhanced cleavage on the N-terminal side of proline is consistent with previous observations for doubly charged ions.[32] The data also suggest that tryptophan can enhance cleavage of the bond on its N-terminal side. Other subtle effects such as the rules regarding mobile versus non-mobile protons for +3 precursors will be the focus of future investigations.

We stress that the method used in this paper is not restricted to any mass spectrometry instrument. As long as a large set of annotated spectra are available, our method can be applied to any proteomics platform. We intend to apply this method to other commonly used MS/MS instruments, e.g. Q-TOF or MALDI/TOF/TOF, as well as more specialized modes of fragmentation, e.g. photodissociation.

The prediction of peptide fragmentation spectra may have other potential applications in protein analysis. It has been shown that the fragmentation patterns may correlate with protein local structures.[33] Including the features of protein local structures in the prediction of peptide fragmentation spectra may result in a potential new approach to protein structure determination.

Acknowledgements

This study was partially funded by the Indiana University Office of the Vice President for Research through a Faculty Research Support grant awarded to RJA, HT and PR.

References

1. Yates JR, 3rd. *Annu Rev Biophys Biomol Struct*. **33** 297 (2004).
2. Russell SA, Old W, Resing KA, Hunter L. *Int Rev Neurobiol*. **61** 127 (2004).

3. Resing KA, Ahn NG. *FEBS Lett.* **579** 885 (2005).
4. Biemann K. *Biomed Environ Mass Spectrom*. **16** 99 (1988).
5. McCormack AL, Jones JL, Wysocki VH. *J Am Soc Mass Spectrom*. 3 859 (1992).
6. Barbacci DC, Russell DH. *J Am Soc Mass Spectrom.* **10** 1038 (1999).
7. Thompson MS, Cui W, Reilly JP. *Angew Chem Int Ed Engl*. **43** 4791 (2004).
8. Zubarev RA, Kelleher NL, McLafferty FW. *J Am Chem Soc*. **120** 3265 (1998).
9. Syka JE, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF. *Proc Natl Acad Sci USA*. **101** 9528 (2004).
10. Sadygov RG, Cociorva D, Yates JR, 3rd. *Nat Methods.* **1** 195 (2004).
11. Yates JR, 3rd, Eng JK, McCormack AL, Schieltz D. *Anal Chem*. **67** 1426 (1995).
12. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. *Electrophoresis*. **20** 3551 (1999).
13. Zhang N, Aebersold R, Schwikowski B. *Proteomics*. **2** 1406 (2002).
14. Moore RE, Young MK, Lee TD. *J Am Soc Mass Spectrom*. **13** 378 (2002).
15. MacCoss MJ, Wu CC, Yates JR, 3rd. *Anal Chem*. **74** 5593 (2002).
16. Biemann K, Martin SA. *Mass Spectrom Rev*. **6** 1 (1987).
17. McCormack AL, Somogyi A, Dongre AR, Wysocki VH. *Anal Chem*. **65** 2859 (1993).
18. Wysocki VH, Tsaprailis G, Smith LL, Breci LA. *J Mass Spectrom*. **35** 1399 (2000).
19. Zhang Z. *Anal Chem*. **76** 3908 (2004).
20. Zhang Z. *Anal Chem* (2005) Electronic publication ahead of print.
21. Elias JE, Gibbons FD, King OD, Roth FP, Gygi SP. *Nat Biotechnol*. **22** 214 (2004).
22. Tabb DL, Smith LL, Breci LA, Wysocki VH, Lin D, Yates JR, 3rd. *Anal Chem*. **75** 1155 (2003).
23. Bairoch A, Apweiler R. *Nucleic Acids Res*. 28 45 (2000).
24. Qian N, Sejnowski TJ. *J Mol Biol*. **202** 865 (1988).
25. Cybenko G. *MCSS, Math Control Signals Syst*. **2** 303 (1989).
26. Riedmiller M, Braun H. *Proc IEEE Internat'l Conf on Neural Networks*. **1** 586 (1993).
27. Saerens M, Latinne P, Decaestecker C. *Neural Comput*. **14** 21 (2002).
28. Kapp EA, Schutz F, Reid GE, Eddes JS, Moritz RL, O'Hair RA, Speed TP, Simpson RJ. *Anal Chem*. **75** 6251 (2003).
29. Quinlan J. *C4.5: programs for machine learning*. San Mateo, CA: Morgan Kaufmann; 1992.
30. Bern M, Goldberg D, McDonald WH, Yates JR, 3rd. *Bioinformatics*. **20** I49 (2004).
31. Baczek T, Bucinski A, Ivanov AR, Kaliszan R. *Anal Chem*. **76** 1726 (2004).
32. Anderson DC, Li W, Payan DG, Noble WS. *J Proteome Res*. **2** 137 (2003).