BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 1 2008, pages 102–109
doi:10.1093/bioinformatics/btm545

Systems biology

Fast and accurate identification of semi-tryptic peptides in shotgun proteomics

Pedro Alves[1], Randy J. Arnold[2,4,*], David E. Clemmer[2,4], Yixue Li[5], James P. Reilly[2,4], Quanhu Sheng[1,4,5], Haixu Tang[1,3,4,*], Zhiyin Xun[2], Rong Zeng[5] and Predrag Radivojac[1,*]

[1]School of Informatics, [2]Department of Chemistry, [3]Department of Biology, Center for Genomics and Bioinformatics, [4]National Center for Glycomics & Glycoproteomics, Indiana University, Bloomington, IN, USA and [5]Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China

Received on August 7, 2007; revised on October 7, 2007; accepted on October 26, 2007

Advance Access publication November 22, 2007

Associate Editor: John Quackenbush

ABSTRACT

Motivation: One of the major problems in shotgun proteomics is the low peptide coverage when analyzing complex protein samples. Identifying more peptides, e.g. non-tryptic peptides, may increase the peptide coverage and improve protein identification and/or quantification that are based on the peptide identification results. Searching for all potential non-tryptic peptides is, however, time consuming for shotgun proteomics data from complex samples, and poses a challenge for a routine data analysis.

Results: We hypothesize that non-tryptic peptides are mainly created from the truncation of regular tryptic peptides before separation. We introduce the notion of truncatability of a tryptic peptide, i.e. the probability of the peptide to be identified in its truncated form, and build a predictor to estimate a peptide's truncatability from its sequence. We show that our predictions achieve useful accuracy, with the area under the ROC curve from 76% to 87%, and can be used to filter the sequence database for identifying truncated peptides. After filtering, only a limited number of tryptic peptides with the highest truncatability are retained for non-tryptic peptide searching. By applying this method to identification of semi-tryptic peptides, we show that a significant number of such peptides can be identified within a searching time comparable to that of tryptic peptide identification.

Contact: predrag@indiana.edu; rarnold@indiana.edu; hatang@indiana.edu

1 INTRODUCTION

Due to its high throughput, the shotgun approach has become a dominant strategy in proteomics. Protein mixtures are treated by trypsin digestion typically followed by reversed-phase liquid chromatography tandem mass spectrometry (RP-LC/MS/MS) (Aebersold and Mann, 2003; Resing and Ahn, 2005; Russell *et al*., 2004; Yates, 2004). The MS/MS spectra obtained from the mass spectrometer are often searched against a protein database by a computer program, e.g. Mascot (Perkins *et al*., 1999) or Sequest (Yates *et al*., 1995). In this peptide identification process, usually only those peptides that follow the rigorous trypsin cleavage rules, i.e. cleavages after a basic residue (arginine or lysine) except when followed by a proline, are considered. We note that missed trypsin cleavage sites and variable post-translational modifications are often included in searches and increase the size of the search, however, these types of searches are not discussed here. It has been shown that only ~10–15% of all tryptic peptides from a given protein sample can be identified with typically 50% of the protein identifications based on a single tryptic peptide (States *et al*., 2006) and that the intrinsic chemical properties of a tryptic peptide have an effect on its probability of being observed in a shotgun proteomics experiment (Lu *et al*., 2007; Mallick *et al*., 2007; Tang *et al*., 2006). Therefore, identifying more peptides, e.g. non-tryptic peptides, preferably at low computational costs would increase the confidence of the proteins identified from the tryptic peptides and potentially increase the overall number of protein identifications.

It is commonly known that trypsin is a specific protease, but it was not well understood how specific trypsin was when treating a protein mixture until a recent work by Olsen *et al*. (2004). They used the high mass accuracy of a linear ion-trap-FTICR mass spectrometer to exclude precursor ions with less than 1 p.p.m. mass accuracy from consideration. In these experiments, spectra that would have otherwise been assigned to non-tryptic peptides were not able to meet this criterion. This work provides evidence in support of the searches for fully tryptic peptides only. Nevertheless, many shotgun proteomics experiments rely only on highly sensitive but lower mass accuracy ion trap instruments. Due to their fast scan rates and the ability to accumulate precursor ion species that would not produce well-resolved, high mass accuracy precursor signals such as in a linear ion-trap-FTICR instrument, these instruments can select and fragment low-abundance non-tryptic peptides. Others have, in fact, shown that non-tryptic peptides are readily identified in proteomics experiments (Tsur *et al*., 2005), and demonstrated the relationship between digestion solvents and existence of non-tryptic peptides (Strader *et al*., 2006).

To estimate how many non-tryptic peptides we can identify in a typical shotgun proteomics experiment, we acquired two

*To whom correspondence should be addressed.

sets (A and B, see Methods section) of MS/MS spectra from two different synthetic mixtures of standard proteins using two different MS/MS instruments (linear ion-trap versus LTQ-Orbitrap) in two independent proteomics labs. We observed among many of the MS/MS spectra that remain unassigned in a routine tryptic peptide search would be identified if the search were performed to include non-tryptic peptides, also referred to as truncated (tryptic) peptides. Furthermore, a majority (>90%) of these non-tryptic peptides are semi-tryptic peptides, which are truncated from one end (either N-terminal or C-terminal) of the tryptic peptide, thus preserving one trypsin cleavage site. Since these samples are made by mixing standard proteins, these truncated peptides are likely formed due to chemical phenomena in the experimental procedures, and are not necessarily the result of proteases in the sample or other biological processes.

The effect of tryptic peptide truncation is demonstrated in Figure 1. In this example, the truncation occurs at the C-terminus of the peptide, so that y-ions are shifted in m/z across the four peptides while b-ions are not. The y-ions labeled y_6, y_9 and y_{11} in part A correspond to fragmentation at the locations of the dashes in YLEFI-SD-AII-HVLHSK with a single charge retained on the C-terminal fragment. Fragmentation between the same residues results in ions y_4, y_7, and y_9 in part B and ions y_3, y_6 (present but not labeled), and y_8 in part C. For the fully tryptic peptide, these fragment ions are among the most intense in the spectrum. For the first two truncated peptides (loss of SK followed by loss of H) the corresponding ions are rather strong. However, in the tandem mass spectrum of the final truncated peptide (further loss of VL) the corresponding peptides are not observed. The varying LC retention times for these peptides (see caption of Fig. 1) suggest that these species are present in the proteolytic digest sample, and not created in the electrospray ion source of the mass spectrometer. The strong signal-to-noise ratio of these MS/MS spectra along with the LC-MS data provides convincing evidence that peptide truncation is an observable phenomenon in shotgun proteomics experiments.

Even though many truncated peptides may be identified in a shotgun proteomics experiment, it is inefficient to apply a conventional database search to identify them, because there are many more potential truncated peptides than tryptic peptides. For instance, one single tryptic peptide of length l can result in $2\cdot(l-1)$ semi-tryptic peptides, and $l\cdot(l-1)/2$ truncated peptides. Hence, even the semi-tryptic search may take at least 10 times more time than the tryptic peptide search. One way to address this issue and speed up the truncated peptide searching is to utilize peptide sequence tags that are generated by the *de novo* sequencing tools (Frank and Pevzner, 2005) to filter non-tryptic peptide sequences in the database (Frank *et al.*, 2005). This approach, however, performs well only for high quality MS/MS spectra, from which good sequence tags can be generated.

In this article, we adopt a different approach for database filtering. We first show that truncation of tryptic peptides is not a uniformly random process—some tryptic peptides are more likely to be truncated than others. We then hypothesize that this non-uniform probability is due to the different chemical properties of tryptic peptides that likely affect the stability of

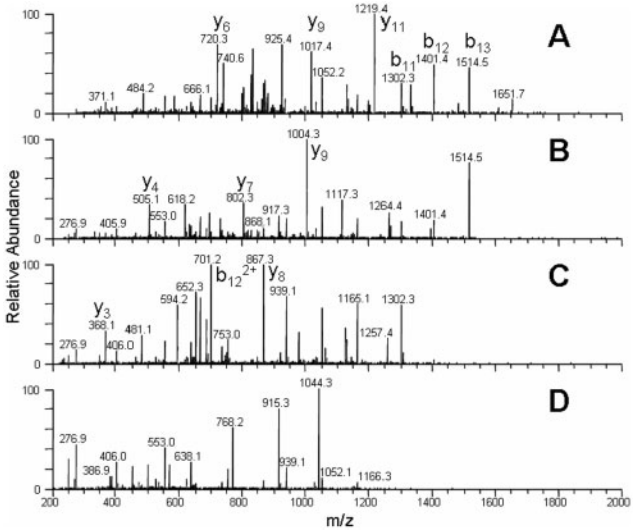

Fig. 1. Ion trap tandem mass spectra for (**A**) the tryptic peptide YLEFISDAIIHVLHSK for horse myoglobin and truncated versions of the same peptide, (**B**) YLEFISDAIIHVLH, (**C**) YLEFISDAIIHVL and (**D**) YLEFISDAIIH. Selected sequence-specific ions are labeled, although all strong peaks in each spectrum can be assigned as sequence fragments. The LC/MS data corresponding to the four peptides are listed as (*peptide*, *mass*, *m/z observed*, *Mascot score*, *LC retention time*, *LC-MS peak area*): (YLEFISDAIIHVLHSK; 1884.01; 943.66; 96; 41.89; 3.54×10^6), (YLEFISDAIIHVLH; 1668.89; 836.00; 98; 42.60; $4.68\ 10^6$), (YLEFISDAIIHVL; 1531.83; 767.76; 90; 44.54; $3.29\ 10^6$), (YLEFISDAIIH; 1319.68; 661.14; 66; 38.82; $8.75\ 10^5$).

peptides in solution. Based on this observation, we propose to predict the truncatability of a tryptic peptide, i.e. the probability of the peptide to be identified in a truncated form, and build a predictor to estimate a peptide's truncatability from its sequence. We show that our prediction achieved high accuracy and can be used to filter the sequence database for identifying truncated peptides. After filtering, only a limited number of tryptic peptides with high predicted truncatability are retained for a truncated peptide search. Since semi-tryptic peptides comprise the majority of non-tryptic peptides in a proteome experiment, we applied this method to accelerating semi-tryptic searches. Our results show that a significant number of semi-tryptic peptides can be identified using computational search times comparable to those needed for conventional tryptic peptide searches.

2 TRUNCATABILITY OF TRYPTIC PEPTIDES

The truncation of a tryptic peptide can be viewed as the loss of one or more amino acids from the N- and/or C-terminus of the fully tryptic peptide. In addition to the biological mechanisms inside a living cell, e.g. proteolytic activities, peptide truncation may be caused by various chemical mechanisms during sample preparation, handling and storage. Even though the process of in-source decay in the mass spectrometer is well understood, we believe that truncation is mainly not facilitated in this way, since we consistently observe that truncated and full tryptic

peptides elute at different retention times in liquid chromatography.

Given a tryptic peptide, we define its *truncatability* as the probability that this peptide is observed in any truncated form in a standard proteomics experiment (e.g. where quantities of all proteins are similar). The requirement of a standard proteomics experiment enables us to better understand physicochemical properties of truncated peptides by eliminating the influence of protein quantities. In this article, we focus on the most popular platforms, the LC coupled to ion-trap (linear or 3D) mass spectrometers. We view the truncatability of a tryptic peptide as an intrinsic property of a peptide for a given set of experimental conditions, and hence expect that it can be predicted from the peptide sequence. We also emphasize that we are particularly interested in the *extremely truncatable peptides*, i.e. the tryptic peptides that are not observed as fully tryptic peptides, but only in their truncated form(s). We rely on these peptides to improve the peptide coverage when analyzing shotgun proteomics data, but note that other truncated peptides can also be useful, e.g. in protein quantification studies based on spectral counts.

3 METHODS

3.1 Data sets

Here we utilized three data sets of MS/MS spectra. Data sets A and B were acquired from synthetic protein mixtures and data set C was acquired from a real proteome sample. Both synthetic samples (A and B) contained proteins that were mixed at similar concentrations. These samples were suitable for studying and learning truncatability of the peptides. Data set C was suitable for studying effects of predicted truncatability in a biological setting. Data sets A and C were previously used by Tang *et al.* (2006) and are described here for the self-containment of this study.

3.1.1 Data set A Data set A contained 12 model proteins mixed at equimolar quantities. The sample was reduced with dithiothreitol (DTT), alkylated with iodoacetamide (IAM) and digested with trypsin at 37°C for 18 h. After acidifying the sample, peptides were loaded onto a 15 mm by 100 μm i.d. trapping column packed with 5 μm BioBasic 18 particles with 300 Å pores (Thermo Hypersil-Keystone, San Jose, CA, USA). Peptides were separated using a 30-min reversed-phase LC gradient from 3% to 40% acetonitrile at 250 nl/min (Eksigent Technologies, Livermore, CA, USA) on a 15 cm, 75 μm i.d. capillary column pulled to a small (~10 μ) tip and packed in-house with 5 μm C18 coated particles (Betasil C18, Thermo Hypersil-Keystone, San Jose, CA, USA). As peptides eluted from the column, they were electrosprayed into the source of a Thermo Electron (San Jose, CA, USA) LTQ linear ion trap mass spectrometer and analyzed by mass spectrometry and tandem mass spectrometry.

3.1.2 Data set B Data set B contained 18 proteins mixed at similar concentrations and was digested with trypsin after treatment similar to that for data set A. The samples were loaded onto the trap column with a 3 μl/min flow rate after the split, and then the reversed-phase gradient was from 2% to 40% mobile phase B in 90 min at 150 μl/min flow rate before the split and 2 μl/min after the split. A linear ion trap/Orbitrap (LTQ-Orbitrap) hybrid mass spectrometer (ThermoFinnigan, San Jose, CA, USA) equipped with an ESI microspray source was used for MS/MS experiments. The mass spectrometer was set so that one full MS scan was acquired in the Orbitrap parallel to three MS/MS scans in the LTQ linear ion trap on the three most intense ions from the full MS spectrum. The resolving power of the Orbitrap mass analyzer was set at 60 000 for the precursor ion scans (at m/z 400).

3.1.3 Data set C Data set C was generated using a complex proteome sample from *Drosophila melanogaster*. Drosophila genotype: elav-GAL4 (Stock number: Bloomington/458) flies were grown for one day and decapitated. Heads were collected on dry ice and stored at −80°C. Proteins were extracted, reduced with DTT, alkylated with IAM and digested with TPCK-treated trypsin. Tryptic peptides were isolated by C18 solid-phase extraction, vacuumed to dryness and stored at −80°C until future use. Peptides were separated by nano-flow reversed-phase liquid chromatography [15 cm × 75 μm i.d. fused silica capillary column pulled to a fine tip and packed with 5 μm, 100 Å amino-terminated C18 packing material (Michrom Bioresources, Auburn, CA, USA), eluted with a gradient from 5% to 45% acetonitrile at 250 nl/min]. Eluting peptides were electrosprayed directly into the source of a Thermo Finnigan LCQ Deca XP ion trap mass spectrometer and analyzed by MS (m/z 250–1500) and data-dependent MS/MS on the three most intense ions.

Tandem mass spectra were searched against protein sequences for the 12 or 18 known proteins (data sets A and B) or all proteins from *D.melanogaster* (data set C) using Mascot for peptide identification. Searches were performed with fixed modification of carbamidomethyl cysteine and variable modifications of protein N-terminal acetylation and methionine oxidation selected and a maximum of one missed trypsin cleavage site. Searches for the non-tryptic peptides in the same data sets were performed using the same modification settings and specifying semi-tryptic cleavage.

3.2 Learning peptide truncatability

We used neural networks to learn peptide truncatability. All peptides identified solely in truncated forms comprised the set of positive examples, while peptides identified as tryptic, regardless of their truncated forms being identified, together with the peptides that were not identified comprised the set of negative examples. For data sets A and B, we used an experience-based default Mascot score threshold of 25 to determine positive identifications, while in the biological sample C we used reversed proteome of *D.melanogaster* for the control of the false discovery rate (5%) by searching both forward and reverse databases together.

3.3 Data representation

Similar to our previous approach (Tang *et al.*, 2006), each tryptic peptide was encoded into a vector form based on its amino acid content and various physicochemical and predicted properties derived from amino acid sequence of the peptide itself and the neighboring peptides within the parent protein. We encoded amino acid compositions, N- and C-terminal residues, and properties such as charge and aromatic content as well as hydrophobic moment (Eisenberg *et al.*, 1984), flexibility by Vihinen *et al.* (1994), B-factor prediction (Radivojac *et al.*, 2004) and disorder prediction (Obradovic *et al.*, 2003; Romero *et al.*, 2001; Vucetic *et al.*, 2003). All predicted properties were encoded as averages within the peptide itself as well as ±5, ±10 and ±15 residues away from the N- and C-terminal residues. N- and C-terminal residues of the peptides were encoded as binary variables. The total number of features was 175.

3.4 Model training

An ensemble of 30 feed-forward neural networks was trained with the final output being an average of individual members. Before network training, we performed feature selection based on the t-test with thresholds from {1, 0.1, 0.01}, where the P-value of 1 corresponds to

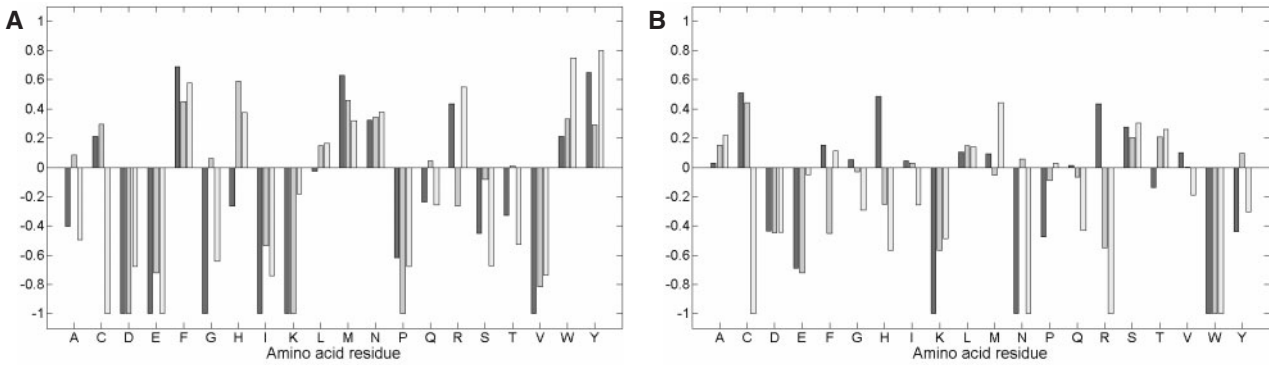

Fig. 2. Amino acid preference at the truncation sites in three data sets used in this study. Black bars—data set A, dark grey bars—data set B; light gray bars—data set C. Each bar is calculated as $(f_t(a)—f(a))/(f_t(a)+f(a))$, where $f_t(a)$ is the relative frequency of amino acid a at (**A**) N-terminal or (**B**) C-terminal side of all truncated sites and $f(a)$ is the relative frequency of amino acid a in all peptides observed as truncated in data sets A, B or C. By definition, instances where there were no truncations for a certain residue result in a value of −1.

retaining all features. The *t*-test was performed after splitting each feature individually into two samples based on the class labels. Then, after the *Z*-score normalization, we performed the principal component analysis retaining 95% of the variance in the sample due to the fact that a large number of features were correlated. Each neural network contained 1, 2 or 4 hidden neurons and was trained using the resilient propagation algorithm (Riedmiller and Braun, 1993). To prevent overfitting, all parameters were selected on the validation data (20% of the training set) and only the final accuracies after the automated parameter selection were reported.

3.5 Model evaluation

The accuracy of the model was estimated using out-of-sample testing where all proteins from one data set were used for training and proteins from the other data set were used for testing. For example, data set A was used for training (20% of set A used for validation and parameter selection) and the final model only was evaluated on data set B. We estimated the balanced-sample accuracy (*accuracy*), i.e. an average between true positive and true negative rates, and the area under the ROC curve (*AUC*).

3.6 Pipeline for identifying semi-tryptic peptides

The prediction of truncatable peptides is carried out off line on the tryptic peptides in the database. Thus, sorting of all the peptides in the database according to their truncatabilities represents a one-time fixed computational cost. A subsequent proteomic search involves only a pre-selected fraction of these peptides (e.g. top 25%) for truncation using the semi-tryptic search option. Note that since the size of the effective peptide database is increased after including all semi-truncated forms of the most truncatable tryptic peptides, the threshold for peptide identifications is expected to be higher than that of a regular tryptic search to maintain the same false discovery rate. For the biological data set described here, e.g. the Mascot score threshold increased from 25 to 32 to maintain 5% false discovery rate. In a slightly different scenario, the proteins that contain at least one identified peptide (tryptic or non-tryptic) can be further searched with a semi-enzyme option to increase their peptide coverage. We call this procedure a two-step identification, where in the first step we identify readily truncated peptides from a list of tryptic peptides (sorted according to a decreased truncatability), and in the second step we perform a semi-tryptic search on the peptides from the proteins identified in the first step (or more precisely, all proteins hit by the peptides identified by the truncatability-enhanced tryptic search).

4 RESULTS

4.1 Data sets

Mascot searches for tryptic peptides resulted in the identification of 164 (114 tryptic, 31 tryptic and truncated and 19 extremely truncated), 134 (62, 29, 43), and 788 (645, 45, 98) peptides for data sets A, B and C, respectively, each with a false discovery rate of 5%. Spectra identified as tryptic peptides could not also be assigned as non-tryptic peptides. Note that for the biological sample (data set C), the sampling spends a smaller proportion of time on less concentrated ions, resulting in the smaller proportion of the identified non-tryptic peptides.

4.2 Analysis of truncated sites

We first analyzed amino acid biases of the identified truncation sites for all three data sets (Fig. 2). In Figure 2A, the amino acid preferences at the N-terminal side follow an interesting trend: there is a significant enrichment of aromatic (F, W, Y) and a significant depletion of acidic (D, E) and hydrophobic residues such as G, P, V and I. While enhancement of truncation with large hydrophobic residues on the N-terminal side is suggestive of chymotryptic activity, we note that only 58.8%, 21.7% and 48.0% of the truncation sites have F, W, Y or M on the N-terminal side observed in data sets A, B and C, respectively, indicating that a large number of truncations cannot be explained simply by the activity of chymotrypsin or chymotrypsin-like activity of trypsin. Moreover, none of the peptides identified in these samples could be assigned to chymotrypsin, further suggesting that the protease itself was not present in the sample in any significant amount. In comparison to the N-terminal side, there is a much larger variability on the C-terminal side of the truncation site (Fig. 2B). Although these data sets were not large enough for complete understanding of the truncation mechanisms, it appears that the possibility of a peptide bond being broken may be influenced more by the residue on its N-terminal side.

4.3 Feature analysis

To improve our understanding of the influence of various properties on peptide truncatability, we used the *t*-test to rank

Table 1. Top 10 features estimated using the *t*-test on merged data sets A and B

Fature	Window	*P*-value	Correlation	Reference
Vihinen *et al.* flexibility	±5	1.7×10^{-8}	−	Vihinen *et al.* (1994)
Hydrophobic moment (angle 120°)	±15	2.6×10^{-8}	−	Eisenberg *et al.* (1984)
VL2-V disorder predictor	±15	2.1×10^{-7}	−	Vucetic *et al.* (2003)
B-factor prediction	±15	3.6×10^{-7}	−	Radivojac *et al.* (2004)
VLXT disorder predictor	±15	4.3×10^{-7}	−	Romero *et al.* (2001)
VL2 disorder predictor	±15	4.6×10^{-7}	−	Vucetic *et al.* (2003)
Peptide length	N/A	1.0×10^{-6}	+	N/A
Hydrophobic moment (angle 100°)	±15	1.0×10^{-6}	−	Eisenberg *et al.* (1984)
Peptide mass	N/A	2.0×10^{-6}	+	N/A
Hydrophobic moment (angle 160°)	±15	3.0×10^{-6}	−	Eisenberg *et al.* (1984)

Features of the same type, but averaged over flanking regions of different sizes, are presented only for the best performing window.

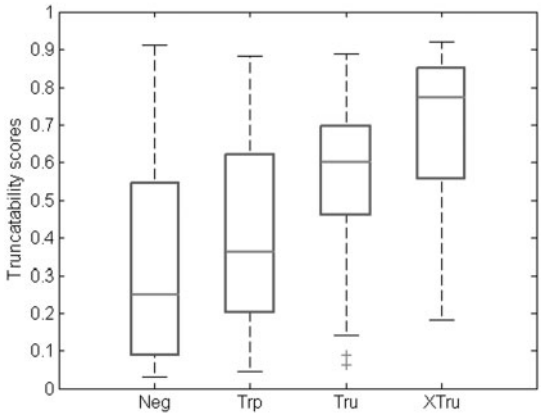

Fig. 3. Box plot of four groups of peptides: (i) Neg—non-identified peptides, (ii) Trp—peptides identified as tryptic only; (iii) Tru—peptides identified as both tryptic and truncated and (iv) XTru—extremely truncated peptides. The number of peptides contained in each group is |Neg| = 171, |Trp| = 62, |Tru| = 29, |XTru| = 43.

individual features from the combined data sets A and B (Table 1). Most significant features indicate that peptides with high local flexibility as well as peptides with high hydrophobic moments are negatively correlated with peptide truncatability. In addition, there exists a positive correlation between the peptide length and mass and the identification of extremely truncated peptides. The mean lengths for the truncated peptides (i.e. their full tryptic versions) in data sets A and B were 19.6 and 17.9, respectively. These peptides were on average longer than identified tryptic peptides in both data sets (12.4 versus 11.4). Data set C contained several very long peptides (111, 106, 98, etc. residues) that resulted in their identification in multiple truncated forms. These long tryptic peptides are extremely truncatable peptides by nature, since they can only be detected in truncated forms. The average length of the truncated peptides in data set C was 28.6, while the average length of the identified tryptic peptides was 15.7. The features shown in Table 1 provide good initial insights into peptide truncation, however, we note that additional experiments are needed to fully understand its chemical basis.

4.4 Truncatability prediction

A predictor trained on data set A and tested on data set B reached accuracy of 67.7% and *AUC* of 76.0%, while a predictor trained on data set B and tested on data set A reached accuracy of 73.6% and *AUC* of 80.0%. Data set C contained proteins at different abundances and was not suitable for training purposes. However, both data sets A and B proved to be good training sets for the evaluation on the biological sample C, reaching *accuracy*/*AUC* of 78.3%/87.4% and 77.9%/86.3%, respectively.

Figure 3 shows a box plot of the scores of all peptides in data set B, separated into four groups: (i) Neg—non-identified peptides, (ii) Trp—peptides identified as tryptic only; (iii) Tru—peptides identified as both tryptic and truncated and (iv) XTru—extremely truncated peptides, i.e. peptides identified only as truncated. The neural network for predicting peptide truncatability was trained using data set A only. Figure 3 shows increasingly larger scores between the four groups of peptides with the extremely truncated peptides (XTru) having the highest scores. This indicates that the highest truncatability scores are most likely to result in identification of extremely truncated peptides.

4.5 Using truncatability in proteomics searches

We used the truncatability predictor trained from the combined data sets A and B to prioritize the search for the semi-tryptic peptides in data set C. We evaluated the trade-offs between the number of searched peptides and the number of identified gained peptides. In Figure 4, we plot the fraction of gained peptides as a function of the fraction of searched tryptic peptides that can result in identification of semi-tryptic peptides in data set C. The fraction of the gained peptides was calculated by first searching the entire *D.melanogaster* database using the semi-trypsin option, which roughly provided us with all identifiable semi-tryptic peptides. Both forward and the reverse sequences were searched at the same time in order to adjust the acceptance thresholds of spectrum-to-peptide matches to the increased database size. A Mascot threshold of 32 was used to achieve a false identification rate of 5%. It is important to note that this threshold is rather conservative for our proposed approach, since in practice only a fraction of peptides will be searched using the semi-trypsin option.

We separately evaluated an algorithm where top *n*% of the tryptic peptides were used for the semi-enzyme search and compared it with the two-step process (see Methods section). In both cases, ~50% of all peptides needed to be searched in

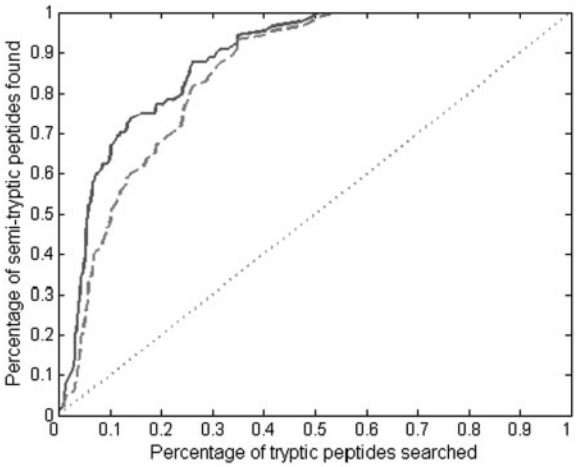

Fig. 4. The fraction of gained peptides (unique extremely truncated peptides) as a function of the fraction of searched tryptic peptides. The fraction of gained peptides was calculated when the number of gained peptides was divided by the total possible number of gained peptides, obtained by the semi-enzyme search. Dashed (green) curve represents database filtering based on predicted peptide truncatability; solid (blue) curve represents filtering based on the two step process, while the dotted (red) line represents a hypothetical baseline case in which peptides are selected randomly.

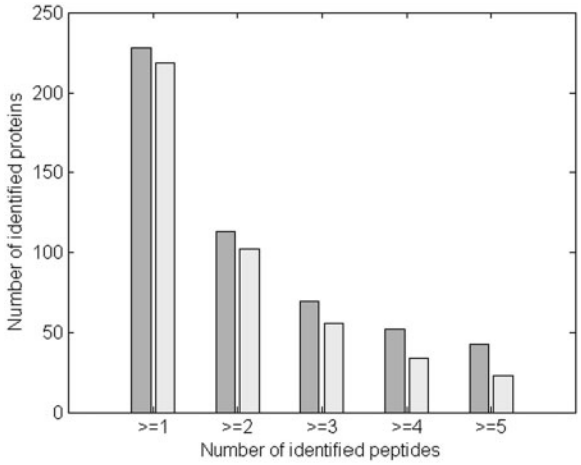

Fig. 5. The number of proteins as a function of the cumulative number of identified peptides in a proteomics experiment. Top 25% of the most truncatable peptides in data set C were searched for non-tryptic peptides. Dark gray bars represent the proposed approach with the prediction of truncated peptides; light gray bars represent the search for tryptic peptides only. Bootstrapping (on the level of identified peptides) with 1000 rounds indicates that the difference between the two distributions is significant.

order to identify all 98 unique extremely truncated semi-tryptic peptides (*gained peptides*). On the other hand, with only the top 10% of the tryptic peptides in the entire proteome, 50% of the semi-tryptic peptides could be identified. Given that in the biological sample proteins could be found at various concentrations and that proteins containing the most semi-tryptic peptides may not even be present in the sample, this result is very promising. As previously indicated, note that if a fraction of total peptides is searched, the acceptance threshold can be further reduced, thus enabling the identification of semi-tryptic peptides that would not have been identified using the no-enzyme or semi-enzyme search on the entire database at the same false discovery rate. When indeed only the top 50% of the most truncatable peptides were selected for the semi-tryptic search, we identified 109 extremely truncated peptides as compared to 98 in the full search (increase of 11%) with the Mascot threshold reduced to 25 (for false discovery rate of 5%) from 32. This number was obtained as an average when 50% of the reverse database was randomly selected as decoy 100 times.

Possible identifications of new proteins and the increased coverage of already identified proteins, especially those identified by a single peptide, can not only improve the confidence of protein identifications, but also help to disambiguate between paralogs in large gene families. Figure 5 shows the distribution of the identified proteins for a given number of peptide hits when the top 25% of the most truncatable peptides (out of 760 617 peptides in total) were used to identify semi-tryptic peptides. It can be observed that the number of proteins in data set C containing two peptides or more increased from 102 to 113, i.e. 11%. At the same time, there was a significantly higher increase in coverage of proteins identified by more than two, three and four peptides, i.e. 25%, 53% and 87%, respectively (Fig. 5). Ten additional proteins were identified by at least one peptide by including semi-tryptic peptides.

4.6 Truncatability versus detectability

In our previous work, we proposed a concept of peptide detectability as the probability that a peptide at standard concentration will be identified in a standard proteomics experiment (Tang *et al*., 2006). We note here that the peptide truncatability is more or less related to detectability, because we can only observe a truncated peptide if it is detectable. However, it is obvious that the truncatability and detectability of a peptide are two different properties. On the one hand, we would classify a peptide as truncatable only if any of its truncated forms is detected. Hence, the highly truncatable peptides would be expected to be detectable, unless the truncations significantly change their detectabilities. On the other hand, peptides that are highly detectable would not necessarily be expected to be highly truncatable. This reasoning is supported by the correlation coefficient of 0.353, 0.611 and 0.597 between the two properties for data sets A, B and C, respectively. The correlation between truncatability and detectability can also be observed through the similar properties of the truncatable and detectable peptides summarized in Table 1 of this study and table 3 from Tang *et al*. (2006). Peptides that are highly hydrophilic (i.e. those having high flexibility and disorder scores) are likely to be washed out in the RP trap and thus have decreased detectability (Galea *et al*., 2006; Tang *et al*., 2006). Those same peptides, on the other hand, are also likely to produce truncatable peptides whose detectability will not be very high. Therefore, it appears that the most important features for finding detectable truncated peptides are highly dependent on peptide detectability. However, the use of

detectability scores, instead of truncatability, to prioritize peptides for searching semi-tryptic peptides resulted in a decreased performance by roughly 50% (data not shown) suggesting that the two properties are indeed different.

5 DISCUSSION

We used trypsin-digested proteins from two synthetic mixtures and a biological sample to demonstrate that non-tryptic peptides, which we refer to as truncated peptides, appear in solution and are readily identified in shotgun proteomics experiments. These truncated tryptic peptides have distinguishable sequence characteristics that can be learned by a machine-learning approach. Knowledge of these characteristics enables the prediction of peptide truncatability that can filter out a large fraction of tryptic peptides before performing a time consuming exhaustive no-enzyme or semi-enzyme search for non-tryptic peptides. For the biological sample, we show that searching only the one quarter most truncatable tryptic peptides (i.e. a saving of $\sim$75% semi-enzyme search time) in all observed proteins allows for $\sim$80% of all identifiable semi-tryptic peptides to be identified (Fig. 4). These additional identified semi-tryptic peptides increase the sequence coverage and confidence in protein assignments. We anticipate that further (more dramatic) time saving can be gained by extending our current approach to enable the prediction of the truncation sites within the truncatable peptides. The non-random distribution of amino acid residues at the N-terminal side of truncation sites (Fig. 2) indicates that these sites are likely predictable. In that case, the number of semi-tryptic peptides that needs to be searched can be reduced to a small constant for each truncatable peptide. Consequently, it will require only comparable time for non-tryptic and tryptic peptide searches.

While it may be expected that truncated peptides are more readily observed in synthetic protein mixture samples than in biological samples due to reduced complexity, we expect that biological samples will have high abundance proteins whose truncated peptides will be readily detected. The results for sample C (Fig. 5) demonstrate this effect, showing that the relative increase in identified peptides when multiple (3 or more) peptides are identified is larger than when only one or two peptides are identified for a protein. Even though the identification of these truncated peptides may not help to identify more low abundance proteins, the increase in peptide coverage can provide useful information for protein inference (Alves *et al*., 2007; Zhang *et al*., 2007) and label-free quantification (Tang *et al*., 2006). Figure 2 supports the argument that the mechanism of peptide truncation, although not yet thoroughly understood, is consistent across both synthetic and biological samples.

It may be argued that since a considerable fraction (in our study $\sim$50%) of non-tryptic peptides are resulted from the chymotrypsin-like cleavages, i.e. after amino acid residues F, W, Y or M (Bender and Kezdy, 1965), one can apply a trypsin/chymotrypsin search, which considers peptide cleavage at all trypsin-specific and chymotrypsin-specific sites. In our analysis, however, even when we allow four missed cleavage sites in this type of search, in which even a much larger peptide database is searched against than using our approach, we can identify only 64% (score $\geq$32) of the semi-tryptic peptides that can be identified (note that only a subset of the identified semi-tryptic peptides are gained peptides) from the entire database, even though 84% (score $\geq$25) of the fully tryptic peptides from a trypsin search are found. This indicates that there are a number of semi-tryptic peptides resulting either from truncation mechanisms other than chymotrypsin-like cleavages or that would require expanding the trypsin/chymotrypsin search even further. Thus, the ideal approach to identification of semi-tryptic peptides appears to be one where all potential semi-tryptic forms from truncatable peptides are considered.

The prediction of truncatable peptides is also useful in studying the proteolysis mechanisms using shotgun proteomics. Identification of semi-tryptic peptides resulted from *in vivo* cleavages can be applied to many important biological problems, e.g. the determination of signal peptides, and the mapping of protein degradation pathways. It is, however, important to differentiate the peptide truncations resulting from *in vivo* biological processes from those caused by chemical phenomena during sample preparation. Our analysis of the synthetic protein mixtures suggests the existence of the latter cause for peptide truncation. It will be useful to predict chemically truncated peptides (and their truncation sites), and exclude them from consideration in biological truncation analysis.

While the training sets used in this study are relatively small, and much larger proteomics data sets from biological sources are available, it is important that biological sources of truncation be eliminated or well characterized so that chemical effects can be studied. The availability of more complex standard protein mixtures would enhance such studies and should be pursued. Further investigation is also needed in order to elucidate the chemical mechanism(s) of peptide truncation.

Previously, we have shown that the estimated peptide detectability can be used to improve protein inference and label-free protein quantification (Alves *et al*., 2007; Tang *et al*., 2006). Even though our prediction of peptide detectability is largely consistent with the peptide identification results in shotgun proteomics, i.e. the peptides with high detectabilities tend to be observed more often than the ones with low detectabilities, there are nevertheless some peptides with high detectabilities that are still not identified, referred to as missed peptides (Alves *et al*., 2007). We observed that the missed peptides often have higher predicted truncatabilities than the non-identified peptides (data not shown). Therefore, these peptides may be missed despite their high detectability because they are extremely truncatable. In the future, we plan to apply the truncatability prediction to improve our protein inference and quantification methods.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their insightful comments. The authors acknowledge the support of the National Cancer Institute, grant # 1 U24 CA126480-01 to F. Regnier, H.T., D.E.C., P.R. *et al.* H.T., R.J.A., D.E.C. and J.P.R. acknowledge NIH/NCRR grant # 5P41RR018942. JPR acknowledges NSF award # CHE-0518234.

Conflict of Interest: none declared.

REFERENCES

Aebersold,R. and Mann,M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207.

Alves,P. *et al*. (2007) Advancements in protein identification from shotgun proteomics using predicted peptide detectability. *Pac. Symp. Biocomput*., **12**, 409–420.

Bender,M.L. and Kezdy,J. (1965) Mechanism of action of proteolytic enzymes. *Annu. Rev. Biochem*., **34**, 49–76.

Eisenberg,D. *et al*. (1984) The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl Acad. Sci. USA*, **81**, 140–144.

Frank,A. and Pevzner,P. (2005) PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal. Chem*., **77**, 964–973.

Frank,A. *et al*. (2005) Peptide sequence tags for fast database search in mass-spectrometry. *J. Proteome Res*., **4**, 1287–1295.

Galea,C.A. *et al*. (2006) Proteomic studies of the intrinsically unstructured mammalian proteome. *J. Proteome Res*., **5**, 2839–2848.

Lu,P. *et al*. (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol*., **25**, 117–124.

Mallick,P. *et al*. (2007) Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol*., **25**, 125–131.

Obradovic,Z. *et al*. (2003) Predicting intrinsic disorder from amino acid sequence. *Proteins*, **53**, 566–572.

Olsen,J.V. *et al*. (2004) Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell Proteomics*, **3**, 608–614.

Perkins,D.N. *et al*. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.

Radivojac,P. *et al*. (2004) Protein flexibility and intrinsic disorder. *Protein Sci*., **13**, 71–80.

Resing,K.A. and Ahn,N.G. (2005) Proteomics strategies for protein identification. *FEBS Lett*., **579**, 885–889.

Riedmiller,M. and Braun,H. (1993) A direct adaptive method for faster backpropagation learning: the RPROP algorithm. *Proc. IEEE Int. Conf. Neural Netw*., **1**, 586–591.

Romero,P. *et al*. (2001) Sequence complexity of disordered protein. *Proteins*, **42**, 38–48.

Russell,S.A. *et al*. (2004) Proteomic informatics. *Int. Rev. Neurobiol*., **61**, 127–157.

States,D.J. *et al*. (2006) Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. *Nat. Biotechnol*., **24**, 333–338.

Strader,M.B. *et al*. (2006) Efficient and specific trypsin digestion of microgram to nanogram quantities of proteins in organic-aqueous solvent systems. *Anal. Chem*., **78**, 125–134.

Tang,H. *et al*. (2006) A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinformatics*, **22**, e481–e488.

Tsur,D. *et al*. (2005) Identification of post-translational modifications by blind search of mass spectra. *Nat. Biotechnol*., **23**, 1562–1567.

Vihinen,M. *et al*. (1994) Accuracy of protein flexibility predictions. *Proteins*, **19**, 141–149.

Vucetic,S. *et al*. (2003) Flavors of protein disorder. *Proteins*, **52**, 573–584.

Yates,J.R., III (2004) Mass spectral analysis in proteomics. *Annu. Rev. Biophys. Biomol. Struct*., **33**, 297–316.

Yates,J.R., III *et al*. (1995) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem*., **67**, 1426–1436.

Zhang,B. *et al*. (2007) Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J. Proteome Res*., **6**, 3549–3557.