BIOINFORMATICS

Vol. 22 no. 14 2006, pages e481–e488
doi:10.1093/bioinformatics/btl237

A computational approach toward label-free protein quantification using predicted peptide detectability

Haixu Tang[1,2,3], Randy J. Arnold[3,4], Pedro Alves[1], Zhiyin Xun[4], David E. Clemmer[3,4], Milos V. Novotny[3,4], James P. Reilly[3,4] and Predrag Radivojac[1,*]

[1]School of Informatics, Indiana University, Bloomington, IN, USA, [2]Center for Genomics and Bioinformatics, Indiana University, Bloomington, IN, USA, [3]National Center for Glycomics and Glycoproteomics, Indiana University, Bloomington, IN, USA and [4]Department of Chemistry, Indiana University, Bloomington, IN, USA

ABSTRACT

Summary: We propose here a new concept of peptide detectability which could be an important factor in explaining the relationship between a protein's quantity and the peptides identified from it in a high-throughput proteomics experiment. We define peptide detectability as the probability of observing a peptide in a standard sample analyzed by a standard proteomics routine and argue that it is an intrinsic property of the peptide sequence and neighboring regions in the parent protein. To test this hypothesis we first used publicly available data and data from our own synthetic samples in which quantities of model proteins were controlled. We then applied machine learning approaches to demonstrate that peptide detectability can be predicted from its sequence and the neighboring regions in the parent protein with satisfactory accuracy. The utility of this approach for protein quantification is demonstrated by peptides with higher detectability generally being identified at lower concentrations over those with lower detectability in the synthetic protein mixtures. These results establish a direct link between protein concentration and peptide detectability. We show that for each protein there exists a level of peptide detectability above which peptides are detected and below which peptides are not detected in an experiment. We call this level the minimum acceptable detectability for identified peptides (MDIP) which can be calibrated to predict protein concentration. Triplicate analysis of a biological sample showed that these MDIP values are consistent among the three data sets.

Contact: predrag@indiana.edu

1 INTRODUCTION

Rapid and reliable identification of thousands of peptides from a complex protein mixture sample using liquid chromatography tandem mass spectrometry (LC/MSMS) and other MS related technologies has established the foundation of high throughput proteomics experiments. Quantitative proteomics, i.e. quantifying proteins in a complex sample, or comparing protein abundances across different samples, however, often requires additional experimental strategies. Several labeling techniques applied to various MS instruments including isotopic coded affinity tag (ICAT) (Gygi *et al*., 1999), mass-coded abundance tagging (MCAT) (Cagney and Emili, 2002), stable isotopic labeling (Oda *et al*., 1999) and global internal standard technology (GIST) (Chakraborty and Regnier, 2002), were developed to profile the differential protein expression of two samples. In spite of their success in some quantitative proteomics experiments, these approaches have their own limitations. For example, some of them target one or several specific amino acids (e.g. ICAT targets Cys and MCAT targets Lys) and thus are limited to those proteins/peptides containing the amino acid that is modified by the reagent. A more important limitation of these approaches is that they all require performing a proper chemical reaction prior to the proteomics analysis. In addition to the expense of chemical reagents involved in this procedure, it remains unclear how the efficiency of these reactions and the protein capturing techniques used in the procedure will affect the quantification of different proteins (Zhang and Regnier, 2002).

Label-free protein quantification approaches attempt to quantify relative protein abundances directly from high-throughput proteomics analyses without applying labeling techniques. Different measures that can be derived from proteomics experiments and presumably correlated to protein abundance were proposed for different MS instruments. For instance, the integration of extracted ion chromatogram (XIC) peaks is thought to be a good measure for LC/MS experiments (Higgs *et al*., 2005) and sophisticated data analysis tools have been proposed to improve its accuracy (Leptos *et al*., 2006). In addition, it has been shown that the spectral count, i.e. the number of times a particular peptide is identified in an experiment, is correlated with the number of protein copies in the sample. Spectral counts have been successfully used to quickly estimate large changes in protein abundance (Pang *et al*., 2002; Gao *et al*., 2003), however the method appears to be significantly less sensitive when the count is relatively small and/or when the difference in protein abundance is 1–2 orders of magnitude (Liu *et al*., 2004; Bonner and Liu, 2006). In summary, there is still lack of systematic testing of the accuracy, robustness and applicability of the label-free protein quantification methods across different MS platforms.

*To whom correspondence should be addressed at School of Informatics, Indiana University, 901 East 10th Street, Bloomington, IN 47408, USA

Here we propose a new approach to label-free protein quantification in high-throughput proteomics experiments based solely on peptide identification, a method that has already been shown to be quite reliable, by learning and applying peptide features to increase the reliability and accuracy of protein quantification. It is commonly observed that the sequence coverage of identified peptides differs from one protein to another in the same proteomics experiment. One may hypothesize that the number of identified peptides or sequence coverage of a protein is highly correlated to its abundance, because the more protein copies in the sample, the higher chance a peptide derived from this protein will be identified (Washburn *et al*., 2001; Ishihama *et al*., 2005; Nesvizhskii and Aebersold, 2005). Although it is intuitively sound, it is not the case in practice. For example, in the analysis of an artificial protein mixture sample, even though twelve proteins were mixed at about the same concentration, the resulting sequence coverage of proteins based on identified tryptic peptides were very different, ranging from almost full coverage to no coverage (Purvine *et al*., 2004). This indicates that the abundance of a protein (or a tryptic peptide from it) is not the only dominant factor that determines whether or not a particular peptide can be observed in a proteomics experiment (Kuster *et al*., 2005).

Several factors related to the nature of the peptides clearly explain the fact that some peptides have higher chances of being missed in the identification than the others even though they are from the same abundant proteins in the sample. Let us use the commonly utilized platform, trypsin digestion coupled with LC/MS analysis, as an example. Peptides with masses smaller than 200 Da and greater than about 6000 Da produce ions (as +1, +2, or +3 ions) that are beyond the m/z range analyzed by the mass spectrometer, typically 200 to 2000 Da, and will not be observed. Other peptides will be so hydrophobic (water-insoluble) that they are not soluble in the LC mobile phase. Still others will be so hydrophilic (water-soluble) that they are not retained by the LC stationary phase in the sample trapping column. In both cases, the peptides will not be ionized for analysis by mass spectrometry. The amino acid composition of some peptides, such as those with multiple acidic residues, may dictate that they do not ionize efficiently in the mass spectrometer ion source. Alternatively, a peptide might ionize well but produce a fragmentation pattern in the MS/MS spectrum that cannot be easily interpreted. Some predicted peptides might never be generated because they exist in a region of the protein's structure that is very stable and thus resistant to proteolysis by trypsin. Finally, each peptide will typically co-elute from the chromatography with other peptides against which it must compete for limited ionizing protons in the electrospray ionization process.

Although these factors are relatively simple and understandable when considered separately, determining the reason for the absence of a peptide is often not straightforward. In fact, it is likely that multiple factors contribute to the overall result—lack of identification. We attempt to learn these 'factors' that govern the likelihood of identifying a peptide by a data driven approach, thus subtract them from the direct correlation between peptide identification and protein quantification, and finally obtain an accurate measure of protein abundance using peptide identification.

This paper is organized as follows. First, we introduce the notion of peptide detectability and discuss its relationship to protein quantification. Next, we show that peptide detectability can be predicted solely from the protein's primary structure with useful accuracy and analyze the sequence features most important for this process. Then, we propose a computational method to quantify a specific protein by using the coverage of identified peptides from a proteomics experiment as well as the predicted peptide detectability. Finally, we demonstrate the robustness of this approach by replicated proteomic analysis on the same sample.

2 PEPTIDE DETECTABILITY

There are four classes of factors that govern the likelihood of observing a peptide in a proteomics experiment: (i) the chemical properties of the peptide (and its parent protein); (ii) the limitation of the peptide identification protocol, including the pre-processing of the sample, the MS instruments and software tools used for mass spectrum analysis; (iii) the abundance of the peptide in the sample; and (iv) the other peptides present in the sample that compete with this peptide in the identification procedure. We define the *detectability* of a peptide as the probability that the peptide will be observed in a standard sample analyzed by a standard proteomics routine. Specifically, we are investigating data from samples treated by trypsin digestion followed by reversed-phase liquid chromatography tandem mass spectrometry in an ion trap and searched against known protein sequences using Mascot (Perkins *et al*., 1999). By standard sample we mean the sample has a fixed number of different proteins (peptides) and they are mixed at the same fixed concentration (e.g. 1 pmol/injection). We stress that, by this definition, peptide detectability is an intrinsic property of a peptide that is determined by its primary sequence as well as its location within the context of the entire protein. Peptides with higher detectabilities have a greater chance of being identified than those with lower detectabilities. As a result, if a peptide with low detectability is identified in a sample, it indicates that this peptide (or the protein this peptide is from) has a high abundance; if a peptide with high detectability is missed (not identified) in a sample, it indicates that this peptide (or the protein this peptide is from) has a low abundance. In addition, a situation in which a peptide with very low detectability is identified while those with higher detectabilities are not, suggests a false positive identification. Therefore, the notion of peptide detectability may be used to establish a direct correlation between peptide identification and protein identification/quantification.

Given a protein, we anticipate that the detectability of all tryptic peptides can be predicted from their sequences. It is, however, important to generate a sample that satisfies the standard conditions we described above, as the learning set for such a prediction. An artificial sample (sample B in Section 3) mixed from 12 model proteins in the similar concentration (1 pmol/microliter) was prepared and analyzed using LC/MS (see Section 5 for details) and the identification results were used as a learning data set for a predictor of peptide detectability in LC/MS experiments. We note that a normal (cellular) proteome sample is not completely suitable for training purposes because proteins in these types of samples have different and unknown abundances.

3 PREDICTION OF PEPTIDE DETECTABILITY

Data sets. We used four groups of data sets of mass spectra in this paper. The first group (data set A) was generated as a standard protein mixture consisting of 12 model proteins and 23 model peptides mixed at similar concentrations from 73 to 713 nM for proteins and from 50 to 1800 nM for peptides (Purvine

Table 1. Composition (fmol per one microliter injection) of six mixtures of 13 model protein chains (12 proteins). This mixture constitutes six data sets: B and B_1–B_5. See Section 5 for detailed description of the sample preparation protocols. MW indicates molecular weight

Protein	Swiss-Prot ID	MW (kDa)	B_1	B_2	B_3	B_4	B_5	B
Serum albumin, bovine	P02769	66.4	3000	300	1000	30	100	1000
Myoglobin, horse	P68082	17.0	3000	300	1000	30	100	1000
Beta-casein, bovine	P02666	23.6	1000	3000	100	300	30	1000
Catalase, bovine	P00432	59.8	1000	3000	100	300	30	1000
Lactoferrin, bovine	P24627	76.1	300	30	3000	100	1000	1000
Lysozyme, chicken	P00698	14.3	300	30	3000	100	1000	1000
Alpha-casein, bovine	P02662	23.0	100	1000	30	3000	300	1000
Pyruvate kinase, rabbit	P11974	57.9	100	1000	30	3000	300	1000
Ovalbumin, chicken	P01012	42.8	30	100	300	1000	3000	1000
DNase I, bovine	P00639	29.1	30	100	300	1000	3000	1000
RNase A, bovine	P61823	13.7	30	100	300	1000	3000	1000
Hemoglobin alpha, human	P69905	15.1	2000	2000	2000	2000	2000	2000
Hemoglobin beta, human	P68871	15.9	2000	2000	2000	2000	2000	2000

Table 2. Summary of the four data sets used in this study. Protein chains with less than 10% sequence coverage were eliminated from all data sets

Data set	Protein chains	Total tryptic peptides	Identified peptides
A	11	346	100
B	13	294	91
C	124	3403	359
D_1–D_3	200	3722	526

et al., 2004). The second group consisted of six data sets (data sets B and B_1–B_5), prepared in our labs, each representing a mixture of the same 13 model protein chains. To mimic a similar peptide competition environment in the LC/MS analysis, we intentionally mixed similar total amounts of protein in each sample as indicated in Table 1. The third group is a data set (data set C) generated from a real rat proteome, as described later. The last group consists of three data sets (data sets D_1–D_3) representing three replicate analyses of the fruit fly head proteome. With the exception of data set C, all samples were reduced and alkylated with iodoacetamide prior to trypsin digestion. The rat samples were digested in the presence of an acid-labile surfactant. All MS experiments were carried out on an ion trap mass spectrometer, either a 3-D ion trap (data sets A, C, and D) or a linear ion trap (data set B). The low m/z cut-off was between 250 and 400, and the high m/z cut-off was between 1500 and 2000 for all experiments.

Due to the large differences in protein concentrations in the whole cell lysates, we included in our analysis and learning procedures only those proteins whose coverage of identified peptides was 10% or higher. In the case of the synthetic sample by Purvine *et al*. (2004), one of the proteins contained only one identified peptide and was also removed from the subsequent analysis. The total number of protein chains, the number of tryptic peptides and the number of identified peptides in each data set are summarized in Table 2.

Machine learning methodology. Given an unseen n-residue long protein sequence $S = s_1 s_2 \cdots s_n$ and a database of peptides already detected by Mascot with high confidence, we construct a model that can approximate the probability of detecting any particular tryptic peptide from S with the same confidence. We denote this probability as $P(score(s_{[i,\ j]}) \geq t \mid S)$, where $s_{[i,\ j]} = s_i s_{i+1} \ldots s_j$ is a residue sequence of a tryptic peptide from S and t is an appropriately selected Mascot threshold (by default 40 in all our experiments). In the case when a Pro residue directly follows a basic residue (Arg or Lys) the peptide was extended until the first accessible Arg/Lys or until the C-terminus. As previously mentioned, in order to reduce the dependency of the detectability on the concentration of the protein in a cell, only proteins with $\geq$10% sequence coverage of the detected peptides were used in our analysis. All peptides whose m/z was outside of the instrument range were eliminated from training and testing as trivial.

Data representation. To enable learning, each input peptide sequence $s_{[i,\ j]}$ was represented by a fixed-length vector of real- or discrete-valued features. Two groups of features were considered: those that depend on $s_{[i,\ j]}$ only and those that also depend on the flanking regions. Thus, an identical peptide observed in the contexts of different sequence neighborhoods will in general have different detectability. The following groups of features were constructed solely from $s_{[i,\ j]}$: (i) amino acid compositions in the peptide; (ii) length of the peptide, i.e. $j - i + 1$; (iii) ion mass m($s_{[i,\ j]}$); (iv) N- and C-terminal residues, s_i and s_j; (v) sequence complexity (Wootton and Federhen, 1996); (vi) physicochemical properties averaged over the entire peptide—aromatic content and hydrophobicity (Kyte and Doolittle, 1982) and (vii) predictions obtained from various bioinformatics tools and averaged over $s_{[i,\ j]}$—namely, protein flexibility predictors (Radivojac *et al*., 2004; Vihinen *et al*., 1994), hydrophobic moment (Eisenberg *et al*., 1984), and predictions of intrinsic disorder (Obradovic *et al*., 2003; Romero *et al*., 2001; Vucetic *et al*., 2003). Since the detectability of the peptide may also be influenced by the neighboring regions, the composite features from (vii) were averaged over the regions of ±5, ±10, and ±15 residues flanking both sides of $s_{[i,\ j]}$. In addition, the residue at position s_{j+1} was also accounted for. Individual amino acids were encoded using orthogonal data representation (Qian and Sejnowski, 1988) while the compositional features were encoded by real values. Overall, the total number of features was 175. A binary class label was finally added

Table 3. Fifteen best features estimated using the t-test on data set B. Features of the same type, but averaged over flanking regions of different sizes, are presented only for the best performing window. Window ±15 indicates that the feature is averaged over $s_{[i-15,\, j+15]}$

Feature	Window	p-value	Correlation	Reference
Vihinen *et al*. flexibility	±15	$3.1 \cdot 10^{-10}$	−	Vihinen *et al*. (1994)
Hydrophobic moment	±15	$6.0 \cdot 10^{-10}$	−	Eisenberg *et al*. (1984)
B-factor prediction	±15	$2.9 \cdot 10^{-9}$	−	Radivojac *et al*. (2004)
VL2 disorder	±15	$1.3 \cdot 10^{-7}$	−	Vucetic *et al*. (2003)
Sequence complexity	0	$1.8 \cdot 10^{-7}$	+	Wootton and Federhen (1996)
VL2V disorder	±15	$3.5 \cdot 10^{-6}$	−	Vucetic *et al*. (2003)
VLXT disorder	±15	$4.1 \cdot 10^{-6}$	−	Romero *et al*. (2001)
VL2S disorder	±15	$4.3 \cdot 10^{-5}$	−	Vucetic *et al*. (2003)
VL3 disorder	±15	$5.5 \cdot 10^{-5}$	−	Obradovic *et al*. (2003)
Composition of Lys	0	$3.3 \cdot 10^{-4}$	−	N/A
Mass/length ratio	0	$1.0 \cdot 10^{-3}$	−	N/A
VL2C disorder	±15	$4.1 \cdot 10^{-3}$	−	Vucetic *et al*. (2003)
Composition of Val	0	$1.6 \cdot 10^{-2}$	+	N/A
Length	0	$1.8 \cdot 10^{-2}$	+	N/A
Composition of Gly	0	$2.1 \cdot 10^{-2}$	+	N/A

to each feature vector; 1 (positive) for a detected peptide and 0 (negative) otherwise.

Model selection. To build predictors we employed ensembles of 30 two-layer feed-forward neural networks trained using the resilient backpropagation algorithm (Riedmiller and Braun, 1993). Due to the asymmetric class sizes and small positive set (detected fragments), each network was trained on a balanced selection of positive and negative examples. Each individual training set contained all the examples from the positive class and the same number of randomly selected negative examples. The network contained 1 output neuron, while the number of hidden neurons h was varied from $h \in \{1, 2, 4\}$. All neurons contained the logistic activation function. Prior to the network training, unpromising features were eliminated using the t-test filter in which features whose p-values were above a given threshold t_{fs} were eliminated. The threshold t_{fs} for feature selection was varied from $t_{fs} \in \{0.01, 0.1, 1\}$. Note that in the case of $t_{fs} = 1$, all features were retained. Finally, correlated features were removed by employing principal component analysis and retaining 95% of the variance. A validation set containing 20% of the training data was used for model selection and overfitting prevention for each of the training sets in the ensemble. Thus, the final prediction was averaged over 30 different models and the single estimated accuracy is reported.

Performance evaluation. The performance of the predictor was evaluated within each data set (A to D) and also across various data sets. In the following, we refer to these two types of performance evaluation as cross-validation and out-of-sample estimation, respectively. In the first case we used a per protein 10-fold cross-validation. The entire set of available proteins D was first split into 10 non-overlapping sets $\{D_i \mid i = 1 \cdots 10\}$. In each step i, dataset $D - D_i$ was used for training while the prediction accuracy was estimated on the test set D_i. The final performance estimates were obtained as averages over all 10 iterations. In the out-of-sample case, we were interested in training and evaluating predictor performance on two independent experiments. In particular, a predictor was trained and optimized on one data set (say, data set A) and then applied and evaluated on all other data sets (say, data sets B, C and D). All twelve combinations were explored.

We measured sensitivity (*sn*)—the fraction of detected peptides correctly predicted, and specificity (*sp*)—the fraction of undetected peptides correctly predicted. Given *sn* and *sp*, the class-balanced accuracy can be calculated as $accuracy = (sn + sp)/2$. In this setup, a predictor always outputting the same class and a predictor outputting uniformly at random would have a balanced-sample accuracy of 50%. In addition to accuracy, we estimated the area under the ROC curve (*AUC*) using the trapezoid rule. Both accuracy and area under the curve are essentially unaffected by the asymmetry in class sizes.

Feature analysis. To gain insights into sequence and physicochemical properties governing peptide detectability, we analyzed features that best discriminate between identified and unidentified peptides. These features were selected using the standard two sample t-test on each feature independently. More precisely, a feature was split into two 1-D samples according to the class label and the hypothesis that these samples were generated according to the same probability distribution was tested. Even though the features may not come from a Gaussian distribution, the t-test is known to be robust to violations of this assumption. In Table 3 we present a ranking according to the increasing p-value of the 15 individually best features obtained on data set B. Nine of these features were based on the overall properties of the peptide including its neighborhood, while the top ranked features based solely on the peptide itself were sequence complexity, its length, the mass/length ratio and presence of Lys, Val, and Gly. Other data sets had similar ordering of the features (data not shown). As a general rule, it appears that peptides within flexible neighborhoods have lower detectability. On the other hand, presence of hydrophobic amino acids (Val, Gly) and peptide length were positively correlated with peptide detectability. Further work is needed toward deeper understanding of these properties.

Prediction accuracy. Predictor evaluation was performed in two steps. In the first step, a 10-fold cross-validation was used to estimate the prediction accuracy on each data set. In the second

Table 4. Results of learning peptide detectability using different training and testing sets. Each field contains balanced sample accuracy (*accuracy*) [%] and the area under the ROC curve (*AUC*) [%] for a particular training/test set combination

accuracy/AUC	Training set A	B	C	D_1–D_3
Test set				
A	75.8/79.7	74.8/80.3	68.0/72.0	63.0/79.2
B	68.3/77.5	65.5/70.0	62.8/69.6	62.7/68.7
C	66.7/74.6	66.8/73.5	75.0/84.0	68.0/78.1
D_1–D_3	78.7/86.5	73.1/79.0	79.9/87.6	86.8/93.0

step, performance evaluation was performed across data sets, as described above. The summary of systematic evaluations is shown in Table 4. Generally, these results strongly support our hypothesis that peptide detectability is influenced by its sequence and flanking regions from the parent protein. Interestingly, the data sets can be grouped into synthetic and whole cell, based on their out-of-sample performance. For example, best out-of-sample accuracy on data sets A and B was achieved when the training sets were B and A, respectively. Training on these synthetic data sets also achieved good performance even on data sets C and D, despite small training sizes. On the other hand, the best out-of-sample performance on data set C was achieved by training on data set D, while the best out-of-sample performance on data set D was achieved by training on C.

It can be observed from Table 4 that prediction accuracies vary between 62.7% and 86.8%, with the mean accuracy of 71.0%, while the area under the curve varied between 68.7% and 93.0%, with the mean of 78.3%. Surprisingly, training on one data set and testing on another did not generally reach similar performance when the two sets were switched. On the one hand, considering the small size of synthetic data sets, such performance could be explained by normal variation. On the other hand, the differences between data sets C and D were large and could be partially explained by the different sample densities in the feature space. In particular, it appears that data sets D_1–D_3 cover only part of the feature space covered by data set C. Thus, while training on C and testing on D_1–D_3 could produce good performance results, the opposite did not hold true. In order to verify this statement we trained a separate classification model to distinguish solely between tryptic peptides from data set C and data set D. A prediction accuracy of 57.4% indicates that there exists a difference between these two samples which can partially explain the inconsistency on the out-of-sample evaluations. In addition to the sequence biases between these two sets, there are also differences in the experimental protocol that could contribute to the discrepancy in performance, e.g. the way in which cysteines were modified in the samples was different for data set C (no modification) and D (reduced and alkylated).

4 PEPTIDE DETECTABILITY AND PROTEIN QUANTIFICATION

In the previous section, we showed that our predictor can approximate detectability of a peptide from its sequence as well as from its context in the complete protein with good prediction accuracy. In this section, we show the results of utilizing the predicted peptide detectability to measure protein abundances in the sample.

Here we analyze samples B_1–B_5 using a predictor trained on sample B in which all chains were similarly abundant. Figure 1a shows the predicted detectabilities of all tryptic peptides from each protein from sample B_1. Peptides from the same protein are shown in the same column, sorted by their detectabilities. Proteins were sorted by their relative abundances (concentrations) in the mixture. The identified peptides are shown as empty squares, while the missed peptides are shown as dashes. It is clear that, for each protein in sample B_1, the identified peptides tend to have higher detectabilities than those not identified. This is consistent to the prediction accuracy results as shown in the last section. For each protein, we can determine its *minimum acceptable detectability of identified peptides* (MDIP), a cutoff value of detectability which maximizes the sum of true positive and true negative rates. If all peptides from a protein are detected, the MDIP of this protein is set to 0, and if none of the peptides from a protein is detected, the MDIP of this protein is set to 1. It can be observed from Figure 1a that the MDIP values, shown as black squares, increase as the protein abundance decreases. This trend is approximated by a solid regression line. Similar results were obtained in the remaining samples B_2–B_5 (data not shown).

We computed the MDIP for each protein in five different synthetic mixtures (B_1–B_5) and show them in Figure 1b. Each column in Figure 1b corresponds to a particular concentration and represents proteins from different experiments. For example, in column 2 the grey diamond and circle represent proteins ALBU_BOVIN and KPYM_RABIT, respectively, both with concentration 1000 fmol. However, ALBU_BOVIN was mixed at this concentration in sample B_3, while KPYM_RABIT was mixed at concentration 1000 fmol in sample B_2 (see Table 1). Similarly to the trend observed in Figure 1a, we can see from Figure 1b a linear relationship between MDIP and protein concentration. Moreover, their relationships are generally similar from one protein to the next.

Figure 2 shows the MDIP for hemoglobin A and hemoglobin B, which were mixed in the same amount in all experiments (Table 1), across different samples. It shows low variation of MDIP, suggesting it is a robust measure of protein abundance.

In the last experiment, we show that MDIP may be used as a measure of protein quantification in high throughput proteomics experiments. Here, we used three replicate data sets (D_1–D_3) to demonstrate the robustness of the protein quantification method we propose. Using the same predictor trained on data set B, we predicted the detectability of all proteins in *D. melanogaster* proteome. In each of the three experiments (D_1–D_3), we computed the MDIP score for each protein. Figure 3 shows the scatter plots of pairwise comparisons of MDIP scores between any two experiments.

5 MASS SPECTRUM ACQUISITION AND ANALYSIS

Data sets B and B_1–B_5. Mixtures of twelve standard proteins (listed in Table 1) were paired or triply-grouped such that the combined molecular weights in each group totaled about 80 to 90 kDa. Samples of each protein were prepared as stock solutions of 60, 20, and 2 micromolar concentration, or 90, 30, and 3 micromolar for

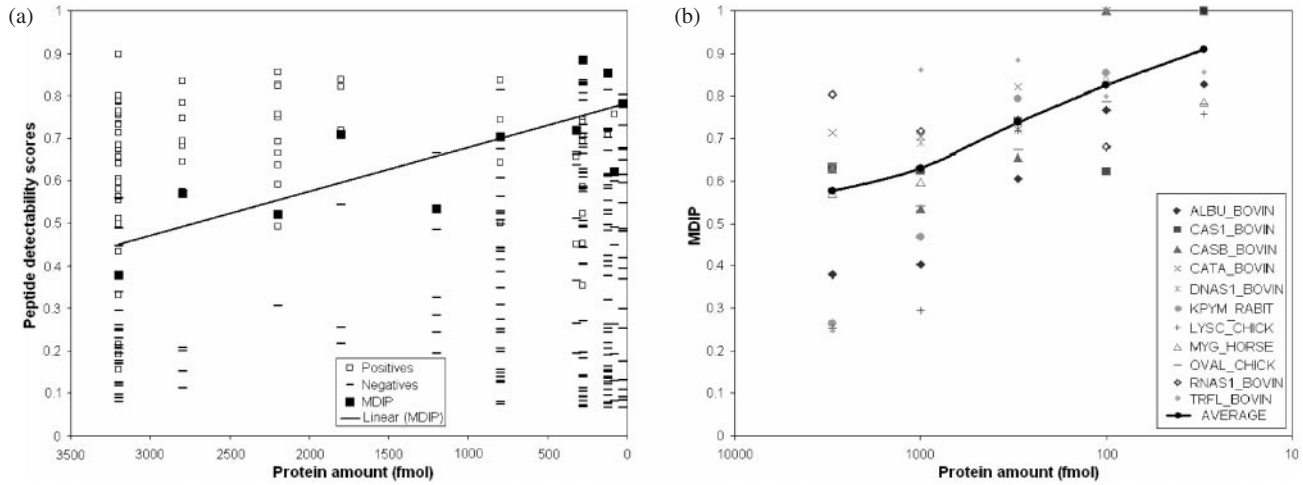

Fig 1. (**a**) Peptide detectability of proteins in sample B_1. Each column displays peptide detectabilities from the same protein. Proteins are sorted according to the decreasing concentration (from left to right), however in order to avoid overlaps, proteins with the same concentration were separated (e.g. columns 1 and 2 correspond to the amount of 3000 fmol). Peptides identified by Mascot are shown as empty squares; peptides not identified are shown as dashes. Minimum acceptable detectability of identified peptides (MDIP) is shown as black squares for each protein. (**b**) MDIP of the proteins from samples B_1–B_5 as a function of protein amount. The columns represent protein amounts and not different experiments. For example, in column 1 RNAS1_BOVIN (top detectability) corresponds to experiment B_5, while CATA_BOVIN (second highest detectability) corresponds to experiment B_2 (see Table1). Both proteins have the abundance of 3000 fmol.

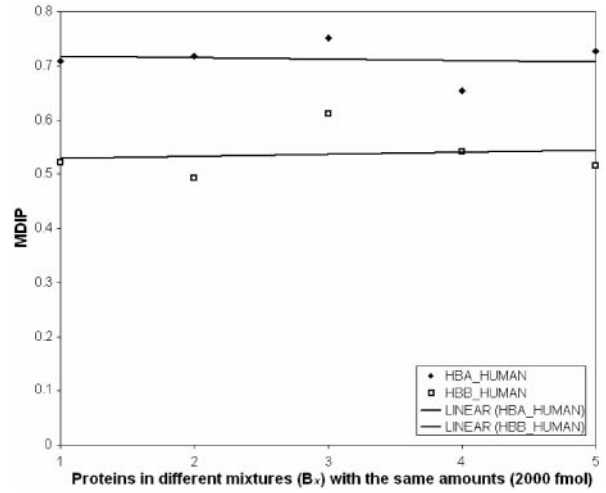

Fig 2. Minimum acceptable detectability of identified peptides (MDIP) of hemoglobin A (HBA_HUMAN, black diamonds) and hemoglobin B (HBB_HUMAN, white squares) in samples B_1–B_5. Each column x in the figure corresponds to a data set B_x.

the triply-grouped samples. Proteins were then mixed in various ratios such that the same molecular weight equivalent was present at 3000, 1000, 300, 100, and 30 fmol per microliter of final digestion solution, combined with buffer, reduced with dithiothreitol (DTT), alkylated with iodoacetamide (IAM), and digested at 37°C for 18 hours. After acidification, samples were loaded onto a 15 mm by 100 micron i.d. trapping column packed with 5-micron BioBasic 18 particles with 300 Å pores (Thermo Hypersil-Keystone, San Jose, CA). Peptides were separated using a 30-minute reversed-phase liquid chromatography gradient from 3% to 40% acetonitrile at 250 nL/min (Eksigent Technologies, Livermore, CA) on a 12 to 15 cm, 75 micron i.d. capillary column pulled to a small (~10 micron) tip and packed in-house with 5 micron C-18 coated particles (Betasil C18, Thermo Hypersil-Keystone, San Jose, CA). As peptides eluted from the column, they were electrosprayed into the source of a Thermo Electron (San Jose, CA) LTQ linear ion trap mass spectrometer and analyzed by mass spectrometry and tandem mass spectrometry. By using dynamic exclusion, the mass spectrometer was limited to acquiring only one tandem mass spectrum for a given parent m/z over a 30-second window.

Data set C. Rat brain regions (amygdala, caudate putamen, frontal cortex, hippocampus, hypothalamus, and nucleus accumbens) 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. Dynamic exclusion was employed to limit acquisition of tandem mass spectra for the same parent m/z over a 60-second window.

Data set D. Drosophila genotype: elav-GAL4 (Stock number: Bloomington/458) flies were harvested and separated according to sex at day 1 of adult life. Flies were cultured on standard cornmeal medium and maintained at 25°C. Flies ($n = 250$) were anesthetized with CO_2, flash frozen and decapitated with shaking in liquid N_2. Heads were collected on dry ice and stored at −80°C. Proteins were extracted using a mortar and pestle in 0.2 M phosphate buffer saline plus 8 M urea plus 0.1 mM phenylmethylsulfonyl fluoride (pH 7.0) solution. Proteins were centrifuged (15700 g at 4°C) for 10 minutes and the supernatant was kept for the determination of protein concentration using Bradford assay. Extracted proteins were reduced with DTT, alkylated with IAM, and digested with TPCK-treated trypsin after diluting the urea to 2 M with

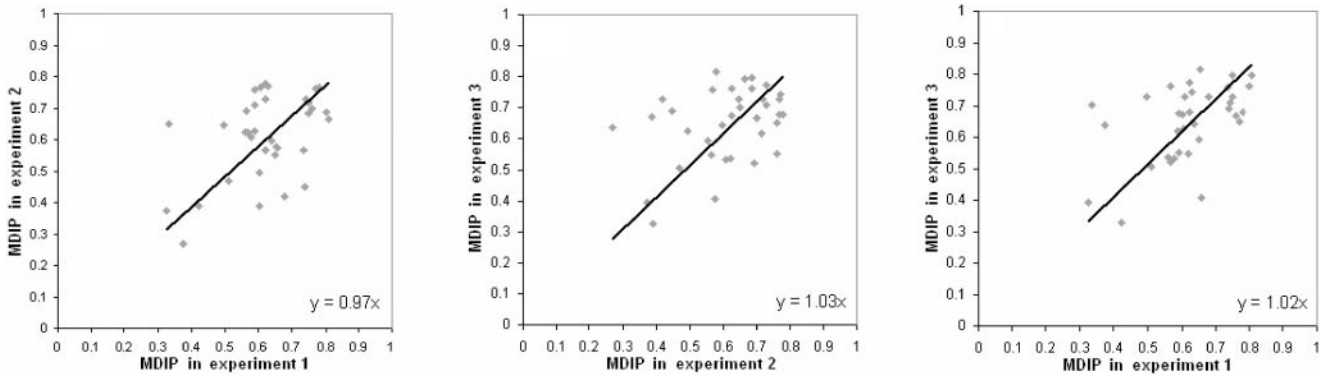

Fig 3. Scatter plots of pairwise comparisons of MDIPs of the identified proteins in samples D_1–D_3. Each dot represents a single protein identified in both experiments.

0.2 M Tris buffer (pH 8.0). Tryptic peptides were isolated by C-18 solid-phase extraction, vacuumed to dryness, and stored at −80°C until future use. Peptides from each SCX fraction 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 C-18 packing material (Michrom Bioresources, Auburn, CA), 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 twelve known proteins (data set B), *R. norvegicus* in the Swiss-Prot database (data set C) or *D. melanogaster* (data set D) using a licensed copy of Mascot (Perkins *et al*., 1999) for peptide identification. Searches were performed with fixed modification of carbamidomethyl cysteine (where appropriate) and 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 training sets including all peptides with Mascot scores of 40 or higher for doubly-charged precursors. Peptides with Mascot scores below 40 were treated as negatives in the training sets.

6 CONCLUSIONS

In this study we propose a new concept of peptide detectability, an intrinsic property of a peptide in the context of its parent protein. This detectability can be used to quantify proteins from the peptide identification results in a standard proteomics experiment. We suggest that peptide detectability can be successfully approximated from its amino acid sequence and neighboring regions of its parent protein. To this goal, we carried out a controlled proteomics experiment in which all protein concentrations were similar to create a ''standard'' data set from which peptide detectability can be learned. In addition to the standard data set B we used other samples to train and evaluate neural-network predictors. Despite small and noisy data sets, these predictors achieved useful cross-validation and out-of-sample accuracies, ranging from 62% to 87%, while the areas under the ROC curves ranged from 69% to 93%.

At this stage, our work is a proof-of-concept study of utilizing the predicted peptide detectability to measure protein abundances in high-throughput proteomics experiments. Further experiments will be necessary in order to precisely determine its sensitivity. It should also be noted that, while demonstrated here as a method to improve quantitative measurements of proteins in proteomics experiments, this approach also offers promise to improve protein identification in cases where only a limited number of peptides are identified.

From the machine learning perspective, we provide only first indications that peptide detectability is predictable from the sequence of its parent protein, thus leaving substantial room for improvement. It is likely that increased data set sizes and variability of samples will contribute to the overall increase in accuracy of detectability prediction, thus somewhat compensating for the class-label noise in the real proteomic samples used in this study. This noise was in part introduced by our simplifying the original problem in which all peptides with Mascot scores <40 were labeled as negative. In addition, we believe that further improvements can be achieved by controlled proteomics experiments in which the informatics approaches proposed here could be properly calibrated.

The results presented here are based on data from a common proteomics analytical platform; nanoflow reversed-phase liquid chromatography coupled by electrospray ionization to tandem mass spectrometry in an ion trap mass spectrometer. Several other analytical methods, such as 2-D liquid chromatography, capillary electrophoresis, MALDI ionization, electron-capture/electron-transfer dissociation, and photoinduced dissociation, as well as alternative proteases are also commonly used in the analysis of complex proteomics samples. Measurements of peptide detectability for analytical platforms based on combinations of these techniques allows for further training, and the potential to determine the most sensitive analytical platform to be used for detection of a specific protein.

ACKNOWLEDGEMENTS

The authors would like to acknowledge William McBride and Wendy Strother-Robinson for help in procuring the rat brain samples, Thomas Kaufman for help in generating the drosophila samples and Narmada Jayasankar for help in data analysis. Dr. Kolker is thanked for providing us with the data generated by Purvine *et al*. (2004). This work is supported in part by the Indiana University Office of the Vice President for Research through a Faculty

Research Support grant awarded to RJA, HT and PR. MVN thanks the Indiana Genomics Initiative (INGEN) of the Indiana University, supported in part by the Lilly Endowment Inc. JPR acknowledges support from the National Science Foundation.

REFERENCES

Bonner, A.J. and Liu, H. (2006) Towards the prediction of protein abundance from tandem mass spectrometry data. *Proceedings of the SIAM International Conference on Data Mining*, **6**, 599–603.

Cagney, G. and Emili, A. (2002) De novo peptide sequencing and quantitative profiling of complex protein mixtures using mass-coded abundance tagging. *Nat. Biotechnol.*, **20**, 163–170.

Chakraborty, A. and Regnier, F. (2002) Global internal standard technology for comparative proteomics. *J. Chromatogr. A*, **949**, 173–184.

Eisenberg, D., Weiss, R.M. and Terwilliger, T.C. (1984) The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. USA*, **81**, 140–144.

Gao, J., Opiteck, G.J., Friedrichs,M., Dongre, A.R. and Hefta, S.A. (2003) Changes in the protein expression of yeast as a function of carbon source. *J. Proteome Res.*, **2**, 643–649.

Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H. and Aebersold, R. (1999) Quantitative analysis of protein mixtures using isotope coded affinity tag. *Nat. Biotechnol.*, **17**, 994–999.

Higgs, R.E., Knierman, M.D., Gelfanova, V., Butler, J.P. and Hale, J.E. (2005) Comprehensive label-free method for the relative quantification of proteins from biological samples. *J. Proteome Res.*, **4**, 1442–1450.

Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J. and Mann, M. (2005) Exponentially modified protein abundance index (emPAI) for estimateion of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell Proteomics*, **4**, 1265–1272.

Kuster, B., Schirle, M., Mallick, P. and Aebersold, R. (2005) Scoring proteomes with proteotypic peptide probes. *Nat. Rev. Mol. Cell Biol.*, **6**, 577–583.

Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.

Leptos, K.C., Sarracino, D.A., Jaffe, J.D., Krastins, B. and Church, G.M. (2006) MapQuant: open-source software for large-scale protein quantification. *Proteomics*, **157**, 1770–1782.

Liu, H., Sadygov, R.G. and Yates, J.R.,3rd (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.*, **76**, 4193–4201.

Nesvizhskii, A.I. and Aebersold, R. (2005) Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell Proteomics*, **4**, 1419–1440.

Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C.J. and Dunker, A.K. (2003) Predicting intrinsic disorder from amino acid sequence. *Proteins*, **53**, 566–572.

Oda,Y., Huang, K., Cross, F.R., Cowburn, D. and Chait, B.T. (1999) Accurate quantitation of protein expression and site-specific phosphorylation. *Proc. Natl. Acad. Sci. USA*, **96**, 6591–6596.

Pang, J.X., Ginanni, N., Dongre, A.R., Hefta, S.A. and Opitek, G.J. (2002) Biomarker discovery in urine by proteomics. *J. Proteome Res.*, **1**, 161–169.

Perkins, D.N., Pappin, D.J., Creasy, D.M. and Cottrell, J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.

Purvine, S., Picone, A.F. and Kolker, E. (2004) Standard mixtures for proteome studies. *OMICS*, **8**, 79–92.

Qian, N. and Sejnowski, T.J. (1988) Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.*, **202**, 865–884.

Radivojac, P., Obradovic, Z., Smith, D.K., Zhu, G., Vucetic, S., Brown, C.J., Lawson, J.D. and Dunker, A.K. (2004) Protein flexibility and intrinsic disorder. *Protein Sci.*, **13**, 71–80.

Riedmiller, M. and Braun, H. (1993) A direct adaptive method for faster backpropagation learning: the RPROP algorithm. *Proceedings of the IEEE Interntional Conference on Neural Networks*, **1**, 586–591.

Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J. and Dunker, A.K. (2001) Sequence complexity of disordered protein. *Proteins*, **42**, 38–48.

Vihinen, M., Torkkila, E. and Riikonen, P. (1994) Accuracy of protein flexibility predictions. *Proteins*, **19**, 141–149.

Vucetic, S., Brown, C.J., Dunker, A.K. and Obradovic, Z. (2003) Flavors of protein disorder. *Proteins*, **52**, 573–584.

Washburn, M.P., Wolters, D. and Yates, J.R.,3rd (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification strategy. *Nat. Biotechnol.*, **19**, 242–247.

Wootton, J.C. and Federhen, S. (1996) Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.*, **266**, 554–571.

Zhang, R. and Regnier, F.J. (2002) Minimizing resolution of isotopically coded peptides in comparative proteomics. *J. Proteome Res.*, **1**, 139–147.