BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 13 2008, pages 1503–1509
doi:10.1093/bioinformatics/btn218

Systems biology

A support vector machine model for the prediction of proteotypic peptides for accurate mass and time proteomics

Bobbie-Jo M. Webb-Robertson[1,*], William R. Cannon[1], Christopher S. Oehmen[1], Anuj R. Shah[2], Vidhya Gurumoorthi[3], Mary S. Lipton[4] and Katrina M. Waters[1]

[1]Computational Biology & Bioinformatics, [2]Scientific Data Management, [3]Applied Computer Science and [4]Biological Separations and Mass Spectrometry, Pacific Northwest National Laboratory

Received on March 18, 2008; revised on April 18, 2008; accepted on April 29, 2008

Advance Access publication May 3, 2008

Associate Editor: John Quackenbush

ABSTRACT

Motivation: The standard approach to identifying peptides based on accurate mass and elution time (AMT) compares profiles obtained from a high resolution mass spectrometer to a database of peptides previously identified from tandem mass spectrometry (MS/MS) studies. It would be advantageous, with respect to both accuracy and cost, to only search for those peptides that are detectable by MS (proteotypic).

Results: We present a support vector machine (SVM) model that uses a simple descriptor space based on 35 properties of amino acid content, charge, hydrophilicity and polarity for the quantitative prediction of proteotypic peptides. Using three independently derived AMT databases (*Shewanella oneidensis*, *Salmonella typhimurium*, *Yersinia pestis*) for training and validation within and across species, the SVM resulted in an average accuracy measure of ~0.8 with a SD of <0.025. Furthermore, we demonstrate that these results are achievable with a small set of 12 variables and can achieve high proteome coverage.

Availability: http://omics.pnl.gov/software/STEPP.php

Contact: bj@pnl.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The explicit goal of proteomics is to identify and quantify all of the proteins present in a cell at a specific moment. However, this is a significant challenge because unlike the genome, the proteins present in a system at any time are dynamic and of varying complexity. Mass spectrometry (MS) offers a high-throughput approach to quantifying the proteome associated with a biological sample and hence has become the primary approach of proteomic analyses. However, this high-throughput capability has led to a computational bottleneck to process and interpret these large spectral datasets.

Smith *et al.* (2002a,b) presented an accurate mass and elution time (AMT) strategy that employed high-resolution MS, specifically Fourier transform–ion cyclotron resonance (FTICR) MS to validate peptide identification made by lower resolution tandem mass spectrometry (MS/MS). The result is a set of peptides that are both unique and detectable based on mass and elution time profiles. This was one of the first approaches to experimentally identify *proteotypic* peptides associated with a specific MS technology. The MS/MS stage of the analysis acquires a list of peptides that are identified using standard database search algorithms, such as SEQUEST (Yates *et al.*, 1995), termed potential mass tags (PMT). The peptide identifications that are then validated using the FTICR in terms of mass and elution time are termed AMTs. Thus the identified peptides collected for the AMT database are a set of peptides associated with the proteins expressed in the samples that are detectable by FTICR. Subsequent identifications only require selection of the peptide from the AMT database based on AMT measurement, resulting in greater sensitivity and increased throughput. This is especially important for complex samples such as plasma (May *et al.*, 2007), because in replicate assays of conventional shotgun MS/MS the lower abundance peptides are less likely to be identified, leading to poorer coverage and reproducibility.

For any MS-proteomics technology, the challenge of experimentally deriving AMT databases for all organisms, or equally deriving proteotypic peptide libraries based solely on experimentation is daunting due to the rapid rate at which new genomes are being sequenced. Nevertheless, significant effort is being placed on cataloging peptides identified by MS/MS over multiple platforms and database search routines as the information becomes available (Craig *et al.*, 2005; Desiere *et al.*, 2006; Jones *et al.*, 2006; Kiebel *et al.*, 2006). These databases have indeed proved useful to evaluate the proteome of organisms for which this data has been amassed by reducing the computational load on database search routines since they must now only search a subset of potential peptide candidates, the proteotypic set (Kuster *et al.*, 2005).

The challenge of deriving these libraries for new organisms remains. However, because there are many known properties associated with the likelihood that a peptide will be identified, such as the number of basic and acidic residues and the hydrophilicity of the peptide, this challenge can be greatly reduced by the computational prediction of proteotypic peptides. Thus, computational approaches to predict proteotypic peptides directly from the primary sequence have been recently reported for shotgun LC–MS/MS and gel-based MS proteomics (Kuster *et al.*, 2005; Mallick *et al.*, 2007; Tang *et al.*, 2006). Different sets of

*To whom correspondence should be addressed.

proteotypic peptides have been reported, depending on the specific experimental set-up. This is expected due to the different strengths of each experimental technique. For example, LC-MS/MS shotgun proteomics studies, such as MudPIT (Delahunty and Yates, 2007) are more likely to identify low-abundance proteins than gel-based proteomics studies (Washburn *et al*., 2001).

Here, we report an approach for the prediction of proteotypic peptides for AMT studies based on simple sequence derived properties using a support vector machine (SVM) classification approach. A unique advantage for identifying proteotypic peptides for AMT studies is that the prediction of the detectable peptides along with accurate elution time prediction (Petritis *et al*., 2006) of these peptides would allow for the *in silico* prediction of an AMT database without the costly and time consuming prior identification of peptides by MS/MS. As a result, accurate prediction of proteotypic peptides for these studies could significantly reduce cost and time.

In order to ensure that the definition of a 'proteotypic' peptide is of high utility for predicting an AMT database, we define proteotypic to be a peptide that has been included in the AMT database at any time that the parent protein is observed, rather than requiring minimal observations of peptides. We demonstrate that the method retains predictability over three complete independent AMT databases collected for the organisms *Shewanella oneidensis*, *Salmonella typhimurium* and *Yersinia pestis*. In addition, feature selection methods demonstrate that only a small number of descriptors are required for prediction of proteotypic peptides for LC–FTICR–MS. Lastly, the results suggest large proteome coverage can be obtained at a small false discovery rate (FDR).

2 METHODS

2.1 AMT proteotypic peptide datasets

Peptide identification from AMT studies for three diverse bacterial species were used for training and validation of the computational method. Descriptions of the sample preparations and experimental protocols for each organism are described elsewhere in detail for *S.oneidensis* (Lipton *et al*., 2006), *S.typhimurium* (Adkins *et al*., 2006) and *Y.pestis* (Hixson *et al*., 2006). The final list of AMT peptides generated for this study were required to have a spatially localized confidence (SLiC) score of 0.7 associated with the probability that the AMT identification is correct (Anderson *et al*., 2006). Further information about the proteomics facility at Pacific Northwest National Laboratory can be found at http://proteomics.emsl.pnl.gov.

To generate the list of positive and negative candidate peptides, only proteins for which at least one peptide had been identified were used from the full protein database of each organism. This is required because experimental protein identification is directly related to the protein expression in the sample. For each protein included in the study, only possible fully tryptic peptides with up to two missed cleavages, at least 6 amino acids in length, and under 6000 Da were generated, resulting in 59 406, 38 681 and 22 671 peptides for *S.oneidensis*, *S.typhimurium* and *Y.pestis*, respectively. From this list of candidates, 25 771, 13 002 and 6889 peptides were positively identified from the AMT databases for each respective organism, summarized in Table 1. The number of peptides and proteins identified are clearly dependent upon the level of study for each organism. For example, *S.oneidensis* has been studied for many years by the Shewanella Federation (http://www.shewanella.org) and thus has the larger number of identified proteins for these three organisms.

2.2 Proteotypic peptide vectorization

To apply a statistical learning algorithm for the classification of proteotypic peptides, each peptide must be represented as an n-dimensional vector of features. The features quantify the characteristics on which the peptides will be classified. We considered 35 features shown in Table 2 that are computable from the sequence and may play a role in the detectability of a peptide. The first set of 15 features describes physico-chemical properties of the peptide and the remaining 20 features capture characteristics of the amino acid composition of the peptide. There is considerable redundancy in some features, which is addressed by the selection of the most relevant features, as described in Section 2.4.

Table 1. Bacterial species protein and AMT dataset information

Organsims	*S.oneidensis*	*S.typhimurium*	*Y.pestis*
Number of possible proteins	4875	4455	4550
Number of proteins identified	3467	2394	1437
Total number of peptides in identified proteins	59 406	38 681	22 671
Observed peptides (positive examples)	25 771	13 002	6998
Unobserved peptides (negative examples)	33 635	25 679	15 673

Table 2. Proteotypic peptide features

Index	Feature
1	Length
2	Molecular weight
3	Number of non-polar hydrophobic residues
4	Number of polar hydrophobic residues
5	Number of uncharged polar hydrophilic residues
6	Number of charged polar hydrophilic residues
7	Number of positively charged polar hydrophilic residues
8	Number of negatively charged polar hydrophilic residues
9	Hydrophobicity—Eisenberg scale (Eisenberg *et al*., 1984)
10	Hydrophilicity—Hopp—Woods scale (Hopp and Woods, 1981)
11	Hydrophobicity—Kyte—Doolittle (Kyte and Doolittle, 1982)
12	Hydropathicity—Roseman scale (Roseman, 1988)
13	Polarity—Grantham scale (Grantham, 1974)
14	Polarity—Zimmerman scale (Zimmerman *et al*., 1968)
15	Bulkiness (Zimmerman *et al*., 1968)
16–35	Amino acid singlet counts

The 25 features described in Table 2 differ considerably in scale. For example, the molecular weight of a peptide may be in the thousands, while counts of each amino acid are small integer values. In order to standardize the contribution of each feature, the features are normalized to have mean zero and unity variance. Prior to SVM classification, a new vector, $\vec{x}$, is transformed based on the normalization factors used in the training phase,

$$\vec{z} = \frac{\vec{x} - \vec{\mu}_T}{\vec{\sigma}_T}, \quad (1)$$

where $\vec{\mu}_T$ represents the vector of mean values associated with each of the 35 features and $\vec{\sigma}_T$ is the associated SDs. The normalization factors are unique to each dataset.

2.3 SVM technique

SVMs are a state-of-the-art statistical learning method for building classification models (Cristianini and Shawe-Taylor, 2000; Vapnik, 1995). The optimization (training) phase of the SVM algorithm finds the best hyperplane to separate the two groups of points in Euclidean space.

This approach allows separation, even under non-linear circumstances, by mapping the data into a high-dimensional feature space that can be linearly separated via a kernel function. This is performed by solving a quadratic programming problem on the data mapped to a kernel function, K. The final classification is a simple linear computation,

$$f(z)=\sum_i \alpha_i K(z,s_i)+b, \quad (2)$$

where $(\vec{\alpha},b)$ defines the separating hyperplane, z is the normalized data as defined in Equation (1), and s_i is the i-th support vector as defined by the training. The score $f(z)$ yields a quantitative measure of how far z is from the separating hyperplane. A common kernel is the polynomial,

$$K(x,y)=(m(\vec{x}\bullet\vec{y})+c)^p, \quad (3)$$

where $\bullet$ represents the dot product. We use the quadratic version with $m=1$, $c=10$ and $p=2$. Training was done using the publicly available GIST software (http://bioinformatics.ubc.ca/gist) on an SGI Altix with 256 GB of shared memory. However, once trained the classification only requires the linear computation in Equation (2).

2.4 Feature selection

Prior to analysis, it is unknown as to which of the features described in Table 2 are the most relevant to proteotypic peptide prediction. For example, several hydrophobicity scales are used to capture multiple chemical perspectives of hydrophobicity, and there is likely to be redundancy. One approach to these problems is recursive feature elimination (RFE; Guyon *et al*., 2002). However, given the iterative nature of the algorithm and the size and density of our training data, performing RFE in this case is computationally prohibitive. An alternative approach, the Fisher Criterion Score (FCS; Bishop, 1995), has been shown to be effective in selecting features for both microarray (Pavlidis *et al*., 2002) and proteomics (Anderson *et al*., 2003) datasets. The FCS is a function that defines the distance between two distributions based on their means and SDs. In the case of the SVM, a distribution is defined by each of the higher order features used in the kernel [Equation (3)]. For a dataset of peptides from both the *positive* and *negative* classes, distribution k has associated means, $\mu_{pos}(k)$ and $\mu_{neg}(k)$, and SDs, $\sigma_{pos}(k)$ and $\sigma_{neg}(k)$, respectively. The FCS for distribution k is defined as:

$$FCS(k)=\frac{\left(\mu_{pos}(k)-\mu_{neg}(k)\right)^2}{\sigma_{pos}(k)-\sigma_{neg}(k)}. \quad (4)$$

A high FCS value indicates that the distributions for the *positive* and *negative* classes of peptides are markedly different when accounting for the variability in the data.

3 RESULTS

The validation of the statistical model was performed within and across three bacterial AMT datasets to evaluate the generalizability of the classifier. The peptides represented in AMT databases are unique, and as such, employing cross-validation (CV) assures that the same peptide is never represented in both the training and testing of the model. This does not preclude the same peptide from being represented in more than one genome of the three bacteria. If two genomes had a large number of peptides in common and one is used to train and the other to validate, then the model would not be independent from the training. However, as seen in Figure 1, the three genomes are highly divergent at the peptide level. The largest fraction of shared peptides is between *S.typhimurium* and *Y.pestis*, approximately 8.6 and 9.9% of the peptides in each genome, respectively. Thus, based on the level of peptide overlap, training and testing on these three different genomes represents independent datasets that one might expect to observe in nature.

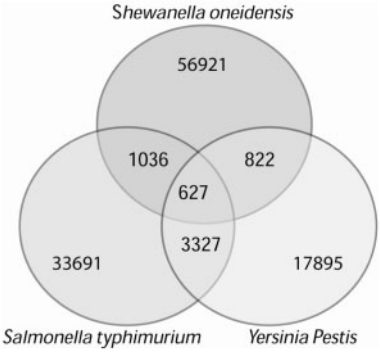

Fig. 1. The number of independent and overlapping peptides for the three datasets of *S.oneidensis*, *S.typhimurium* and *Y.pestis*. Table 1 gives the respective numbers of proteotypic peptides within each genome.

3.1 Validation within and across species

A 3-fold CV study was performed to determine the ability of the method to predict proteotypic peptides for AMT studies within each of the three bacterial datasets. Each dataset was divided equally into three independent sets of roughly equal size with approximately the same ratio of positive and negative candidates for identification. The generalizability of the algorithm is examined by using the classifier developed for one organism to classify the full set of peptides for the other two organisms. The sensitivity and specificity of the method over the peptide space for each organism is measured as a receiver operating characteristic (ROC) curve. The sensitivity measures the number of peptides classified in the positive class (proteotypic) that are ranked above a specific threshold corresponding to a specificity value. The area under the ROC curve (AUC) is a good overall measurement of accuracy, or the ability to correctly classify a peptide on average. A perfect classification method will have an AUC of one; a random binary classifier will have an AUC of ~0.5.

The 3-fold CV of each of the *S.oneidensis*, *S.typhimurium* and *Y.pestis* datasets returned AUCs of ~0.79, ~0.78 and 0.83, respectively—Figure 2. Using a two-tailed signed rank test with a Bonferroni correction (Salzberg, 1997), the ROC curves between *S.oneidensis* and *S.typhimurium* are not statistically different at a P-value of 0.05; however, *Y.pestis* has a ROC curve that is statistically different than the other two at a P-value <0.001. The smaller size of the *Y.pestis* AMT dataset results in less variability at the peptide level compared to the other two datasets and thus is more successful at proteotypic peptide classification using CV. The impact of dataset size can be further tested by evaluating the performance of each of the three classifiers on completely independent data.

For validation across organisms, each classifier is used on the other two datasets; for example, the SVM classifier generated from the *S.oneidensis* dataset is used to classify the peptides for the remaining two organisms, *S.typhimurium* and *Y.pestis*. Figure 2a displays the ROC curves of *S.typhimurium* and *Y.pestis* when classified on the *S.oneidensis* trained SVM. The ROC curve for the intra-species CV for *S.oneidensis* is also included for comparison. Likewise, Figure 2b and c display the cross-species ROC curves for training on *S.typhimurium* and *Y.pestis*, respectively. The resulting AUC values are given in Table 3. For these analyses, the mean AUC is 0.792 with a SD of 0.024, suggesting that despite datasets with clearly different peptide composition (Fig. 1), the method

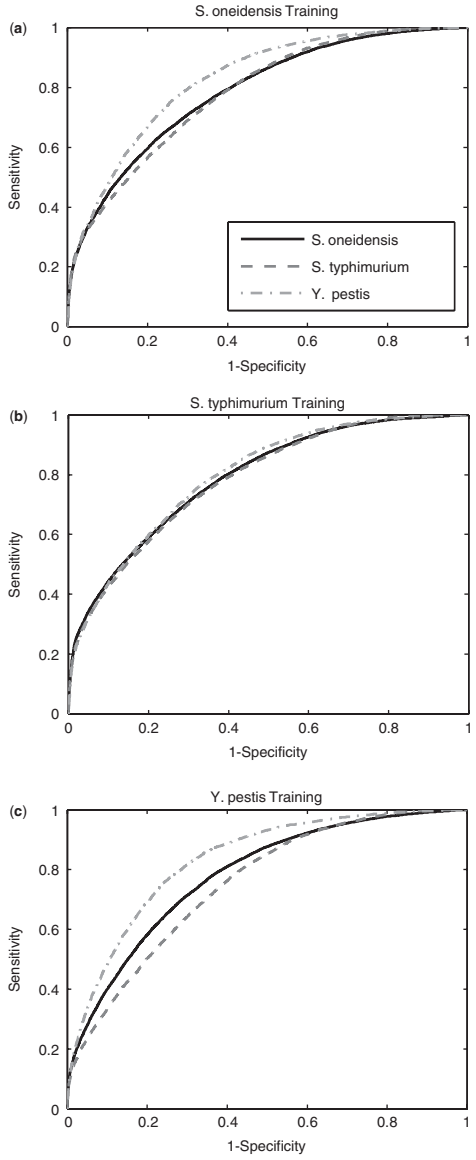

Fig. 2. The ROC curves associated with cross-species CV. (**a**) ROC curves for *S.typhimurium* and *Y.pestis* proteotypic peptide prediction when trained on *S.oneidensis* data. (**b**) ROC curves for *S.oneidensis* and *Y.pestis* proteotypic peptide prediction when trained on *S.typhimurium* data. (**c**) ROC curves for *S.oneidensis* and *S.typhimurium* proteotypic peptide prediction when trained on *Y.pestis* data. Three-fold CV ROCs are included for each organism as well. Table 1 gives the respective numbers of proteotypic peptides within each genome.

has relatively constant accuracy independent of the AMT dataset used for training. Not surprisingly, the smallest classifier, *Y.pestis* (Fig. 2c), returns the most variability in the proteotypic peptide prediction step.

3.2 Feature contributions

Although SVMs are an accurate and robust approach to statistical classification, they come at the cost of reduced explanatory power. Unlike decision trees or exploratory data analysis methods, such as principal component analysis, there are no nodes or latent variables from which to explicitly select and explain relevant features. Thus, we use a Fisher score analysis to evaluate the relevance of features. For each of the features described in Table 2, and for each AMT dataset, we compute the FCS score using Equation (4). The FCS values are used to order the features with respect to their individual ability to separate the proteotypic from non-proteotypic peptides. The FCS and feature orderings are available in Supplementary Tables 1–3. To evaluate the feature contributions across three datasets, the FCSs were averaged for each feature in Table 2. Additionally, it was determined if larger values of the specified feature resulted in either an increased (red) or decreased (blue) capability to separate the proteotypic class. These results are displayed in Figure 3.

Table 3. AUC values for within and across AMT dataset evaluation

Training organism	Testing organism		
	S.oneidensis	*S.typhimurium*	*Y.pestis*
S.oneidensis	0.787	0.783	0.826
S.typhimurium	0.790	0.784	0.797
Y.pestis	0.781	0.752	0.826

There are clear trends in the features of Figure 3 that deserve to be mentioned. First, the amino acids cysteine (feature 17) and lysine (feature 24) are clearly significant discriminators of proteotypic peptides. Cysteine content is a significant feature for classifying peptides that are not proteotypic. Previously, it has been noted that cysteine is underrepresented in detected peptides (Huang *et al.*, 2005). This may be due to the tendency for cysteine to form cross-linked peptides, which are very difficult for current peptide identification algorithms to identify. However, it is not uncommon to reduce disulfide bonds during sample preparation. Thus, it may be that the observability of these peptides has more to do with the use of generic fragmentation models in peptide identification programs than it does with cysteine-containing peptides not surviving sample processing and electrospray ionization. On the other hand, lysine content, a basic amino acid that generally carries a positive charge due to the electrospray ionization process, is a significant feature for classifying proteotypic peptides. These observations seem counter-intuitive at first; however, lysine is at times more likely to result in a peptide carrying an overall positive change than arginine, even though arginine has a greater gas phase basicity. The reason for this is that arginine may form a salt bridge in the gas phase with the carboxylate groups of glutamic acid, aspartic acid and the C-terminus (Price *et al.*, 1997; Schnier *et al.*, 1996), resulting in a net charge of zero for the acid–base pair. However, the true nature of these features in their discriminatory power can only be achieved through carefully designed experiments.

From Figure 3, features 6, 7 and 8 are also all important positive features for identifying proteotypic peptides, and all involve charged, hydrophilic residues. Charge is clearly important for peptides being captured through the electrospray process, and hydrophilicity is important for each step of sample preparation, liquid chromatographic separation and electrospray ionization. Interestingly, hydrophobicity is not a strong feature for classification, although it should be noted that the concept of hydrophobicity is not the opposite of hydrophilicity (Ben-Naim, 1992).

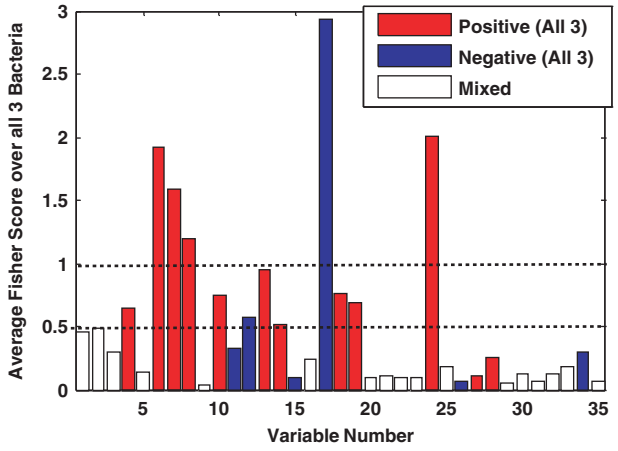

Fig. 3. The average FCS for each of the 35 variables described in Table 2. Red and blue bars indicate features for which the direction of the contribution of the variable was consistent across all three organisms.

There were five variables that had an average FCS of over 1.0. This number increases to 12 for a FCS cutoff of 0.5. Figure 3 is color coded to represent cases in which there is agreement across all three datasets on the discriminatory direction of each variable; red means that the variable has a larger value for the proteotypic peptides and blue is vice versa. The gray bars represent cases in which there is not agreement across the three datasets. In all cases where the average FCS is over 0.5, the class contribution agrees over all three datasets. Given the low FCS for the remaining 23 features these variables likely add little value to the model.

To further evaluate the feature reduction potential, the same analysis as observed in Figure 2, was completed for the dataset with the top 5 variables (FCS > 1.0) and the top 12 variables (FCS > 0.5). To determine if a subset is substantially different from the full set of 35 features, we compute an average AUC and SD overall comparisons, both within and across species. Boxplots of these results showed no outliers. A two-tailed signed rank test with a Bonferroni correction showed that the model with fives variables did not perform as well as the full 35 variables model with a *P*-value >0.05. However, the 12 and 35 variables models were not statistically different from one another at a *P*-value of <0.008. Several other combinations of variables of size less than 12 were also evaluated, but were not as accurate as the 35 variables model at a *P*-value of 0.05.

4 DISCUSSION

Given the peptide challenge associated with MS, due largely to incomplete fragmentation, noisy data and post-translational modifications (PTMs), the computational derivation of proteotypic peptides leads to several questions.

(1) How many true proteotypic peptides are not identified with conventional database searches?

(2) How many peptides are required to build an adequate classification model for proteotypic peptides?

(3) How does peptide detectability affect proteome coverage?

With respect to the first question, a typical MS analysis will leave up to 80% of the spectra without a confident identification to a peptide. There are several potential reasons for this, such as PTMs or partial fragmentation. However, it is also clear that the generic fragmentation models used by most algorithms do not represent reality very well (Webb-Robertson and Cannon, 2007). Thus, the overall accuracy measures of AUCs near 0.8 are not surprising. Methods to predict proteotypic peptides will always have some level of error due to peptides that should be identified because they are present in the experimental spectra, but are not due to improper fragmentation models. It is hypothesized that some fraction of these unidentified spectra could be resolved by sequence specific fragmentation models, if available (Craig *et al.*, 2005; Lam *et al.*, 2007; Yates *et al.*, 1998). Models, such as proposed here and elsewhere (Alves *et al.*, 2007; Mallick *et al.*, 2007; Tang *et al.*, 2006) will need continual updating as peptide database search routines continue to improve and supply more accurate training data.

The second question regarding the number of peptides required to build an adequate classification model for proteotypic peptides is prompted by the diversity of possible peptide sequences. For example, just for a 6mer there are 20^6 unique peptides, yet not all of these will need to be sampled to build a classifier. In preliminary work, we used a small quality control dataset of ~2000 peptides and found very good predictability within the dataset using CV. However, we found little-to-no predictive capability when applied to our organism-level AMT datasets displayed in Figure 2: in all cases the AUC was ~0.5. Thus, clearly for the task of prediction of proteotypic peptides, a large and diverse training set is required. Furthermore, it is necessary to train the dataset on the appropriate data for the required coverage. For example, this demonstration only included fully tryptic peptides and microorganisms. Extension to partial tryptic peptides and higher organisms will likely require the model to be re-trained.

Specifically, in a SVM, a weight is returned for each of the training observations. If the weight of an observation is non-zero, then this observation is an active member of the separating hyperplane. A trivial classification task would have only two support vectors, one for the positive and one for the negative class. As the data complexity increases, the number of active support vectors increases up to a maximum of the number of training examples. In the small quality control dataset, we observed that all of the ~2000 training examples were active support vectors. For the large training datasets used here, the active support vectors correspond to tens of thousands of peptides. In fact, for all three datasets, ~60% of the training vectors are active members of the classifier. The consequence is a classification hyperplane of high dimension, which is not likely to be sufficiently sampled (trained) using a small set of peptides, such as those supplied by small controlled experiments. Thus, large datasets with sufficient peptide diversity are required to achieve predictive power because very little colinearity exists among the training vectors for this problem. Fortunately, given the simplicity of computing the SVM score, Equation (2), even with tens of thousands of support vectors, this requires a few seconds of computation time.

With this in mind, one of the primary benefits of the SVM method described here is the reduction in the search space of the peptide database that the experimental spectra are compared to. By selecting the required fraction of true positives (proteotypic peptides) to be retained in this database, based on the associated threshold a reduced representation of the database can be constructed.

This is a significant computational savings since the computation time would decrease linearly with the peptide search space reduction size. Figure 4a displays the number of proteotypic peptides that would be generated for each of the three AMT databases used in this evaluation at FDR ranging from 0% to 50% (using the *S.oneidensis* trained classifier) with respect to the total number of peptides in each bacterial genome. There is similar agreement on size of the search space between the three organisms, which are of roughly equal genome size.

An additional benefit in this regard is that the SVM prediction provides a rapid means of constructing an AMT database compared to the laborious work of running many MS/MS analyses on samples of unknown protein content. For example, after extensive experimentation on *S.oneidensis* ∼71% of the proteome is represented in the AMT database. However, for *Y.pestis* much less experimentation has been completed and as a result the AMT database only covers ∼32% of the proteome. Figure 4b gives the proteome coverage based on the peptides that are classified as proteotypic at varying FDRs as seen in Figure 4a. Proteome coverage is defined as the fraction of the total number of proteins that have at least one proteotypic peptide associated with them. The curve decreases slightly if one requires at least two proteotypic peptides.

At a FDR of 10%, all three proteomes of the species investigated here have a coverage of nearly 90% using the SVM method on the 12 relevant features identified (Fig. 3). This results in a significant decrease in the search space. For example, the *S.oneidensis* search space is reduced to ∼30% of the original search space and retains ∼88% of the proteome coverage. The proteome coverage improves to nearly 94% at a 25% FDR, which reduces the search space to ∼44% of the original search space. In fact, the predicted proteome coverage is very near the experimentally derived coverage, 50–60% of the proteome, at only a 1% FDR. Although the model can be used to derive an AMT database independent of experimentation, in practice the most successful strategy would likely use the MS/MS phase to validate the computational AMT database. However, since MS/MS would be used as a validation step, the number of sample preparation and runs would be considerably less than deriving an AMT database from scratch. This could have a significant impact on AMT studies in respect to the overall costs in terms of both money and labor.

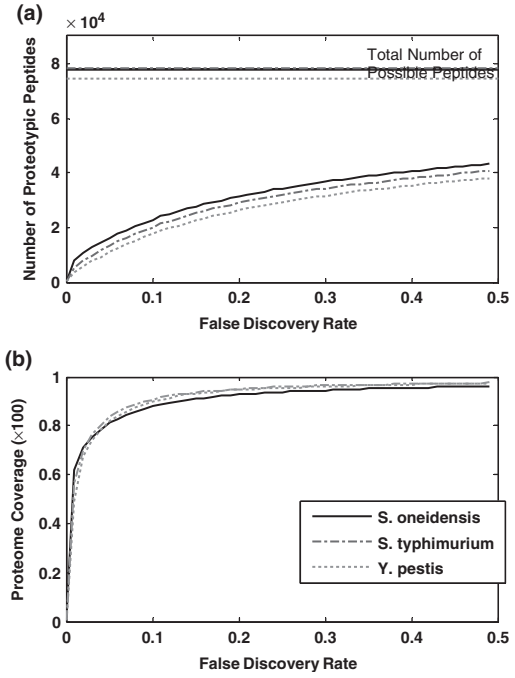

Fig. 4. The predictive (**a**) number of peptides and (**b**) associated proteome coverage based on proteotypic peptides identified for the full genome at FDRs ranging from 0% to 50%.

5 CONCLUSIONS

Peptide identification is one of the fundamental challenges of MS-based proteomics (Lu *et al*., 2007; Webb-Robertson and Cannon, 2007). As samples become more complex and include biological communities, the future of proteomics relies on the ability to accurately perform these identifications at a rate that can keep pace with high-throughput techniques such as AMT studies. The SVM method presented here offers an approach to perform quantitative *in silico* prediction of peptides that are detectable, which when combined with elution time prediction could provide for the rapid creation of an AMT database and reduce the need for extensive MS/MS analyses. Furthermore, the ability to define proteotypic peptides will have a cascading effect that will ultimately yield more accurate statistical prediction of identification, as has been demonstrated for peptide and protein quantification (Alves *et al*., 2007; Tang *et al*., 2006). For example, the statistical assessment of proteotypic peptides clearly demonstrates that the assumptions of equal probability across peptides in protein discriminative models (Nesvizhskii *et al*., 2003) does not capture the true potential of protein identification. Furthermore, as decoy databases are becoming more common to capture the random error of peptide identification, their implementation will be more powerful if true potential proteotypic peptide candidates based on physico-chemical properties are represented and not simply randomly selected peptide sequences.

ACKNOWLEDGEMENTS

We thank the researchers in the laboratory of Dr Richard Smith at PNNL who generated the datasets herein.

Funding: This work was supported through the Laboratory Directed Research and Development at Pacific Northwest National Laboratory (PNNL) and the US Department of Energy (DOE) Office of Advanced Scientific Computing Research under contract No. 47901. PNNL is a multiprogram national laboratory operated by Battelle for the US DOE under contract DE-AC06-76RL0 1830.

Conflict of Interest: none declared.

REFERENCES

Adkins,J.N. *et al*. (2006) Analysis of the Salmonella typhimurium proteome through environmental response toward infectious conditions. *Mol. Cell. Proteomics*, **5**, 1450–1461.

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

Anderson,D.C. *et al*. (2003) A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *J. Proteome Res.*, **2**, 137–146.

Anderson,K.K. *et al*. (2006) Estimating probabilities of peptide database identifications to LC-FTICR-MS observations. *Proteome Sci.*, **4**, 1.

Ben-Naim,A.Y. (1992) *Statistical Thermodynamics for Chemists and Biochemists*. Springer, New York.

Bishop,C. (1995) *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford.

Craig,R. *et al*. (2005) The use of proteotypic peptide libraries for protein identification. *Rapid Commun. Mass Spectrom.*, **19**, 1844–1850.

Cristianini,N. and Shawe-Taylor,J. (2000) *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, Cambridge.

Delahunty,C.M. and Yates,J.R., 3rd (2007) MudPIT: multidimensional protein identification technology. *BioTechniques*, **43**, 563, 565, 567.

Desiere,F. *et al*. (2006) The PeptideAtlas project. *Nucleic Acids Res.*, **34**, D655–D658.

Eisenberg,D. *et al*. (1984) Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.*, **179**, 125–142.

Grantham,R. (1974) Amino acid difference formula to help explain protein evolution. *Science*, **185**, 862–864.

Guyon,I. *et al*. (2002) Gene selection for cancer classification using support vector machines. *Mach. Learn.*, **46**, 389–422.

Hixson,K.K. *et al*. (2006) Biomarker candidate identification in Yersinia pestis using organism-wide semiquantitative proteomics. *J. Proteome Res.*, **5**, 3008–3017.

Hopp,T.P. and Woods,K.R. (1981) Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl Acad. Sci. USA*, **78**, 3824–3828.

Huang,Y. *et al*. (2005) Statistical characterization of the charge state and residue dependence of low-energy CID peptide dissociation patterns. *Anal. Chem.*, **77**, 5800–5813.

Jones,P. *et al*. (2006) PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res.*, **34**, D659–D663.

Kiebel,G.R. *et al*. (2006) PRISM: a data management system for high-throughput proteomics. *Proteomics*, **6**, 1783–1790.

Kuster,B. *et al*. (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.

Lam,H. *et al*. (2007) Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics*, **7**, 655–667.

Lipton,M.S. *et al*. (2006) AMT tag approach to proteomic characterization of Deinococcus radiodurans and Shewanella oneidensis. *Methods Biochem. Anal.*, **49**, 113–134.

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.

May,D. *et al*. (2007) A platform for accurate mass and time analyses of mass spectrometry data. *J. Proteome Res.*, **6**, 2685–2694.

Nesvizhskii,A.I. *et al*. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.*, **75**, 4646–4658.

Pavlidis,P. *et al*. (2002) Learning gene functional classifications from multiple data types. *J. Comput. Biol.*, **9**, 401–411.

Petritis,K. *et al*. (2006) Improved peptide elution time prediction for reversed-phase liquid chromatography-MS by incorporating peptide sequence information. *Anal. Chem.*, **78**, 5026–5039.

Price,W.D. *et al*. (1997) Is arginine a zwitterion in the gas phase? *J. Am. Chem. Soc.*, **119**, 11988–11989.

Roseman,M.A. (1988) Hydrophobicity of the peptide C=O...H-N hydrogen-bonded group. *J. Mol. Biol.*, **201**, 621–623.

Salzberg,S.L. (1997) On comparing classifiers: pitfalls to avoid and recommended approach. *Data Min. Knowl. Disc.*, **1**, 317–327.

Schnier,P.D. *et al*. (1996) Blackbody infrared radiative dissociation of Bradykinin and its analogues: energetics, dynamics, and evidence for salt-bridge structures in the gas phase. *J. Am. Chem. Soc.*, **118**, 7178–7189.

Smith,R.D. *et al*. (2002a) The use of accurate mass tags for high-throughput microbial proteomics. *Omics*, **6**, 61–90.

Smith,R.D. *et al*. (2002b) An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics*, **2**, 513–523.

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

Vapnik,V. (1995) *The Nature of Statistical Learning Theory*. Springer, New York.

Washburn,M.P. *et al*. (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.*, **19**, 242–247.

Webb-Robertson,B.J. and Cannon,W.R. (2007) Current trends in computational inference from mass spectrometry-based proteomics. *Brief. Bioinform.*, **8**, 304–317.

Yates,J.R., 3rd *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.

Yates,J.R., 3rd *et al*. (1998) Method to compare collision-induced dissociation spectra of peptides: potential for library searching and subtractive analysis. *Anal. Chem.*, **70**, 3557–3565.

Zimmerman,J.M. *et al*. (1968) The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.*, **21**, 170–201.