Verification of a Support Vector Machine Model for Predicting Proteotypic Peptides

Jessica Forbes
Carroll College, Helena, MT

Follow this and additional works at: https://scholars.carroll.edu/mathengcompsci_theses

Part of the Applied Mathematics Commons, and the Chemical Engineering Commons

Recommended Citation
https://scholars.carroll.edu/mathengcompsci_theses/42

This Thesis is brought to you for free and open access by the Mathematics, Engineering and Computer Science at Carroll Scholars. It has been accepted for inclusion in Mathematics, Engineering and Computer Science Undergraduate Theses by an authorized administrator of Carroll Scholars. For more information, please contact tkratz@carroll.edu.
Verification of a Support Vector Machine Model for Predicting Proteotypic Peptides

Jessica Forbes

Carroll College, Helena, Montana

April 11, 2010

Prepared in fulfillment of the requirement of graduation with Honors from Carroll College and graduation from the Department of Mathematics, Engineering, and Computer Science.
This thesis for honors recognition has been approved for the Department of Mathematics, Engineering, and Computer Science.

Dr. Holly Zullo – Director

Dr. Kelly Cline – Reader

Dr. Jennifer Geiger – Reader
# Table of Contents

SIGNATURE PAGE ........................................................................................................................................... 2

ACKNOWLEDGEMENTS .................................................................................................................................... 4

ABSTRACT ............................................................................................................................................................ 5

Chapter 1 – INTRODUCTION ............................................................................................................................ 6

Chapter 2 – BACKGROUND .............................................................................................................................. 9

Support Vector Machine (SVM) .......................................................................................................................... 9

Receiver Operating Characteristic (ROC) Curve and Area Under the ROC curve (AUC) ......................... 14

Chapter 3 – MATERIALS AND METHODS ...................................................................................................... 17

Strain selection and AMT proteotypic peptide databases ............................................................................. 17

Protein selection for STEPP .............................................................................................................................. 17

STEPP SVM model evaluation ............................................................................................................................ 18

Chapter 4 – RESULTS ........................................................................................................................................ 19

Chapter 5 – DISCUSSION AND CONCLUSION ............................................................................................. 21

REFERENCES ....................................................................................................................................................... 22
ACKNOWLEDGEMENTS

This work was supported by the U.S. Department of Energy Office of Science through the Science Undergraduate Laboratory Internship (SULI) program and the Laboratory Directed Research and Development program at Pacific Northwest National Laboratory (PNNL). I would like to thank Dr. Bobbie-Jo Webb-Robertson and Dr. Lee Ann McCue for being my SULI mentors. I would like to thank Dr. Holly Zullo for being my Honors Thesis Advisor, and Dr. Jennifer Geiger and Dr. Kelly Cline for being my Honors Thesis Readers. I would also like to thank Kelly Stratton, Karen Wieda, and Matt Monroe from PNNL for their assistance.
ABSTRACT

The current method to match mass spectra from tandem mass spectrometry (MS) to a peptide sequence requires searching a large database of all possible peptides encoded by an organism. However, only a subset of these possible peptides is consistently and repeatedly identified by MS (proteotypic peptides). Matching spectra to this smaller, proteotypic peptide search space increases computational efficiency and improves accuracy of the peptide identification, hence increasing the confidence that a protein has been accurately identified. Currently, it is labor-intensive to build a proteotypic peptide database of experimentally observed peptides; thus computationally deriving such a database is desirable. Webb-Robertson et al. trained a statistical learning algorithm called a support vector machine (SVM) from Yersinia pestis data that computationally classifies a peptide as proteotypic or not proteotypic. Preliminary tests by these authors showed that this SVM accurately predicted proteotypic peptides for two closely related bacterial species – Salmonella typhimurium and Shewanella oneidensis. To test the versatility of the classifier, experimentally generated proteotypic peptide databases from three bacteria more distantly related to Y. pestis, as well as one vertebrate species, were gathered – Pelagibacter ubique, Caulobacter crescentus, Cyanothece, and Mus musculus (mouse). For each of these species, those proteins with at least four experimentally determined proteotypic peptides were extracted and all possible peptides for those proteins were classified with the SVM. The resulting information was analyzed in MatLab, creating a Receiver Operating Characteristic (ROC) curve and associated area under the curve (AUC) value to describe the sensitivity and specificity of the SVM model, where an AUC of 1.0 describes a perfect classifier and a random binary classifier would generate an AUC value of 0.5. The average AUC values for Y. pestis, P. ubique, C. crescentus, Cyanothece, and mouse were 0.8351, 0.7442, 0.7622, 0.7455, and 0.7457 respectively. Therefore, the current SVM classifier accurately predicts proteotypic peptides for diverse bacterial species as well as the mouse. Future research may include retraining SVMs to target a specific protein sample preparation method or species.
Chapter 1 - INTRODUCTION

A key objective of proteomics is the ability to describe the complex and ever-changing protein environment of cells. Unlike static DNA, protein concentrations fluctuate depending on cellular requirements, environmental cues and disease state. For example, protein variability has been important for the detection of biomarkers of disease, by identifying new cellular components and alterations to protein complexes, or organelle composition [1].

Typically in global proteomics studies, proteins are identified via mass spectrometry (MS). Briefly, protein samples are prepared from whole cells, the proteins digested with trypsin, and the resulting peptide mixtures separated by liquid chromatography and analyzed by mass spectrometry. The peptides are identified by matching the mass spectra data with protein sequences. This peptide identification step is complicated by several factors, including: (i) peptide mixtures from whole cells are highly complex and contain peptides of widely varying concentrations, (ii) mass spectrometry technology includes many variations on the ionization technique used and the mass analyzer, (iii) protein preparation procedures vary considerably depending on the sample source (e.g., animal, plant or bacterial cells). These factors affect which peptides will be observed by MS [1]. Proteotypic peptides are those that are consistently and repeatedly detected by MS. Building a database of proteotypic peptides facilitates future studies by allowing focused
searches of mass spectra to these readily observable peptides rather than a
database of all possible peptides. However, building a database by
experimentally identifying proteotypic peptides is extremely labor-intensive.
Therefore, developing a method to computationally predict proteotypic
peptides is an area of great interest in proteomics research.

Methods to predict proteotypic peptides have been developed, including
an artificial neural network [2], a classical pattern discovery and statistical
learning approach [3], a classical machine learning approach [4], and a
support vector machine (SVM) [1]. Webb-Robertson et al. [1] used the
biophysical and biochemical characteristics of peptides and databases of
experimentally determined proteotypic peptides to train an SVM classifier to
identify proteotypic peptides. An SVM is a machine learning algorithm that
is trained to separate linear and non-linear data via a quadratic
programming algorithm, building parallel hyperplanes that separate two
classes of data [5]. Webb-Robertson et al. [1] showed that the SVM Technique
for Evaluating Proteotypic Peptides (STEPP) trained on data from one
bacterial species not only accurately predicted proteotypic peptides from that
same species, but also from other closely related bacterial species. Thus, the
STEPP classifier was validated in a three-fold cross validation on the data
from a single species, as well as validated on completely independent training
and testing datasets from different species.
In this study, the accuracy of proteotypic peptide prediction for data from a more diverse set of species was examined. The SVM trained using experimentally identified proteotypic peptides from *Yersinia pestis* was used to predict proteotypic peptides from three distantly related bacteria – *Pelagibacter ubique, Caulobacter crescentus,* and *Cyanothece,* as well as a vertebrate – *Mus musculus,* to help determine the versatility of the current STEPP classifier, and whether it would be necessary or beneficial to train separate SVM classifiers for different organisms or protein preparation procedures.
Chapter 2 – BACKGROUND

Support Vector Machine (SVM)

A support vector machine (SVM) is a statistical learning algorithm to

distribute objects in categories. For example, an SVM can distinguish

between fraudulent and regular credit card activities. In this case, massive

amounts of data on fraudulent and regular card activity are used to train the

algorithm [5].

We will use the example proposed by William S. Noble of a seminal

study of acute leukemia expression profiles to describe how an SVM works.

The Affymetrix microarrays, applied by Golub et al., yielded mRNA levels for

6,817 human genes. Golub et al. completed this assay on 38 bone marrow

samples, 27 of which were samples from patients with acute lymphoblastic

leukemia (ALL) and 11 samples from patients with acute myeloid leukemia

(AML) [5].

Noble explains that there are four basic concepts which will aid in our

understanding of SVMs: (i) the separating hyperplane, (ii) the maximum-

margin hyperplane, (iii) the soft margin and (iv) the kernel function [5].

For easy interpretation of an SVM, Noble selected data from a previous

study with only two genes – ZXY and MARCKSLI – which were used in an

SVM classifier. Figure 1 displays this data. In the figure, values are

proportional to the intensity of fluorescent on the microarray. Therefore, a
high value on either axis represents a high expression of that gene for the patient. Patients with ALL and AML are indicated on the graph by dark and light points, respectively [5].

**Figure 1** Support vector machines (SVMs) at work. (a) Two-dimensional expression profiles of lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) samples. Each dimension corresponds to the measured mRNA expression level of a given gene. The SVM’s task is to assign a label to the gene expression profile labeled ‘Unknown’. (b) A separating hyperplane. Based upon this hyperplane, the inferred label of the ‘Unknown’ expression profile is ‘ALL’. (c) A hyperplane in one dimension. The hyperplane is shown as a single black point. (d) A hyperplane in three dimensions. (e) Many possible separating hyperplanes. (f) The maximum-margin hyperplane. The three support vectors are circled. (g) A data set containing one error, indicated by arrow. (h) A separating hyperplane with a soft margin. Error is indicated by arrow. (i) A nonseparable one-dimensional data set. (j) Separating previously nonseparable data. (k) A linearly nonseparable two-dimensional data set, which is linearly separable in four dimensions. (l) An SVM that has overfit a twodimensional data set. In a, b, d–h, the expression values are divided by 1,000. (Figure provided by W.S. Nobel) [5]
The separating hyperplane

It is simple for the human eye to see patterns, such as the cluster of AML patients in the upper left portion of Figure 1a and ALL in the lower right portion. It would be easy to state a simple rule that a patient has AML if their \textit{MARCKSLI} level is twice as high as their \textit{ZYX} level. This is seen by drawing a line between the two data clusters as seen in Figure 1b. Therefore it is easy to determine whether a value should be considered ALL or AML by which side of the line the point is found [5].

The maximum-margin hyperplane

To understand the concept of a separating hyperplane, we will consider cases where there are not two genes. For instance, if a study is performed on only one gene, then the ‘space’ for the one-dimensional expression profiles exists on a one-dimensional line, Figure 1c. We would then be able to divide the AML patients from the ALL patients with a single point. In the two-dimensional analysis, the separating factor was a line. If the analysis has a higher dimensional study of more than two genes, the separating factor would be a hyperplane, Figure 1d [5].

When we reexamine Figure 1a we recognize that many separating lines exist between the AML and ALL groups, as seen in Figure 1e. To determine the best classifier, we could choose the line that separates the two
groups and maintains the maximal distance from any one of the given expression profiles, Figure 1f [5].

The soft margin

Thus far, we have assumed that the data can be separated by a straight line. In reality, there will be cases as illustrated in Figure 1g where there is an expression profile associated with the other class. In order to incorporate these occasional 'errors' in the classifier, we implement a 'soft margin' in the SVM classifier. This allows data points to cross the margin of the separating hyperplane without influencing the final result, as shown in Figure 1h. It is possible for the user to specify an allotted number of outliers and how far away these outliers can be from the hyperplane. These restrictions ensure that the SVM classifier does not allow too many misclassifications [5].

The kernel function

In order to understand the idea of a kernel function, Nobel simplified the example even further to an analysis of only one gene. In the case of Figure 1c, the data was easily separated by a single point. On the other hand, Figure 1i demonstrates a nonseparable data set where AML values group near zero and the ALL values have large absolute values. It is not possible to separate the groups by a single point and introducing a soft margin would not solve the problem [5].
Utilizing a kernel function will correct this issue by adding another dimension to the data. The original expression values were simply squared (Figure 1j) and a distinguishing line was drawn to separate the two classes [5]. Essentially, the kernel allows an SVM to perform a ‘two-dimensional’ data analysis on originally one-dimensional data. Typically, a kernel function projects a lower dimensional data space to a higher one [5].

Consider the two-dimensional data set in Figure 1k. Here, the data were projected to a four-dimensional space by squaring the pairs of data. Once in this higher dimension, the classes were easily separated by a line. Nobel could not illustrate a four-dimensional space so the data were projected down to a two-dimensional space and the line is represented by a curve in Figure 1k [5].

Even though it is possible to project every lower dimension to a higher dimension to identify a separating hyperplane, there are consequences to taking data to too high of a dimension. The data in Figure 1k was projected to a higher dimension that essentially overfit the data set in Figure 1l. Here, the result is small clusters that have been specified to the training data too much. This often must be avoided through the use of trial and error. Scientists usually start with a simple kernel and work their way through several generic kernels [5].
Conclusion

An SVM can almost perfectly classify the ALL/AML data set using all 6,817 gene expression measurements. An even larger data set study demonstrated that an SVM actually performs better than other cancer classification methods. Even though Nobel's description focused on cancer classification, SVMs can be utilized in many biological studies [5].

Receiver Operating Characteristic (ROC) Curve and Area Under the ROC curve (AUC)

The Receiver Operating Characteristic (ROC) Curve is a graphic representation of a classifier's accuracy. It is often utilized to select a threshold for the classifier to assure a defined level of sensitivity or specificity of the classifier. The SVM classifier assigns a probability value, of belonging to the positive group (proteotypic) or the negative group (not proteotypic), to each peptide sequence of the search. A probability value is then selected as a threshold to define if a sequence is classified correctly or not, resulting in four categories. The four categories are: (i) true positives (TP) are true proteotypic peptides predicted to be positive, (ii) false positives (FT) are true non-proteotypic peptides predicted to be positive, (iii) true negatives (TN) are true non-proteotypic predicted to be negative, and (iv) false negatives (FN) are true proteotypic peptides predicted to be negative. An ROC curve is a plot of the true positive rate (TPR) versus the false positive rate (FPR), on the y and
$x$ axes respectively. TPR is the ratio of correctly predicted positives to the total actual positives, those actual positives predicted as positive or negative.

$$TPR = \frac{TP}{TP + FN}$$

The TPR is also known as sensitivity. FPR is the ratio of correctly predicted negatives to the total actual negatives; those actual negatives predicted as positive or negative.

$$FPR = \frac{FP}{FP + TN}$$

The FPR is also known as $(1 \cdot \text{specificity})$. Therefore, ROC curves are equivalently labeled as sensitivity vs. $(1 \cdot \text{specificity})$ of the classifier.

As explained earlier, the SVM produces continuous probability values "representing the degree to which class the instance belongs." If this value is greater than a threshold value, it will be placed in the positive class. Therefore, graphing the ROC points for each threshold value produces a curve.

A perfect discrete classifier would assign a point to $(0,1)$ of the ROC space. This demonstrates 100% sensitivity, no FN, and 100% specificity, no FP. A random classifier would assign a point on the line $y = x$, called the line of no-discrimination. The numerical representation of the accuracy of the classifier is indicated by the area under the ROC curve (AUC). Therefore, a perfect classifier would have an AUC value of 1.0 because at all values of the FPR the TPR is 1, essentially creating a box for the ROC curve. The random
classifier is a straight line from (0,0) to (1,1), which creates a lower triangle of the box, which would have an AUC value of 0.5. The closer the AUC value is to 1.0, the more accurate the classifier.
Chapter 3 - MATERIALS AND METHODS

Strain selection and AMT proteotypic peptide databases

For this study, databases of MS observed peptides for three distantly related bacteria – Pelagibacter ubique, Caulobacter crescentus, and Cyanothece – and Mus musculus (mouse) were used to investigate the versatility of the SVM model. These organisms were chosen to reflect broad genetic diversity. Experimentally observed peptide data were obtained from the proteomics facility at Pacific Northwest National Laboratory (http://proteomics.emsl.pnl.gov). A MatLab® script was used to eliminate redundancy in the peptide databases, leaving only unique peptides positively identified by the mass spectrometer. Redundancy occurs because the same peptide sequence can occur in more than one protein. This redundancy must be removed to avoid bias in the calculation of classification accuracy. For example, the P. ubique and C. crescentus data were reduced by approximately 11% and 12%, respectively.

Protein selection for STEPP

For each species, those proteins with greater than four peptides in the unique peptide file were extracted from the file of all proteins encoded in the genome and used to test the Y. pestis STEPP classifier. This increases our confidence in the original identification of those peptides and proteins chosen.
to test the STEPP proteotypic peptide classification. The protein sequences were then randomly divided into three subsets and saved in fasta format.

**STEPP SVM model evaluation**

Each protein fasta file was analyzed using the STEPP program, which provides a list of all the possible peptides from those proteins, and a corresponding score for the likelihood that the peptide is proteotypic. This information was analyzed in MatLab, creating an ROC curve and an AUC value.
Chapter 4 – RESULTS

The validation of the STEPP classifier was completed using three bacteria distantly related to the training organism *Y. pestis*. Previously [1], two closely related bacteria were used to evaluate the algorithm’s ability to accurately classify proteotypic peptides across different species. They reported AUC values for *Shewanella oneidensis, Salmonella typhimurium*, and the source organism, *Y. pestis*, of 0.781, 0.752, and 0.826 – respectively. In this study, the SVM performed well for *Y. pestis* with an average AUC value of 0.8351. The experimentally observed proteotypic peptide database for *Y. pestis* is considerably larger than that available to Webb-Robertson et al. [1], which may explain our improved AUC value (0.826 to 0.8351) for *Y. pestis*.

For validation across a wider genetic range, the classifier was evaluated using the three bacteria *P. ubique, C. crescentus*, and *Cyanothece*, as well as *Mus musculus* (mouse). Each fasta sequence file was reduced to only those proteins with at least four peptides present in the database of unique, experimentally observed proteotypic peptides for that organism. These proteins were randomly distributed into three subsets for use in the STEPP program. Figure 2 displays the ROC curves for these four organisms and Table 1 lists the corresponding AUC values.
Table 1. The AUC values for *Y. pestis*, *P. ubique*, *C. crescentus*, *Cyanothece*, and *M. musculus*.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Set 1 AUC</th>
<th>Set 2 AUC</th>
<th>Set 3 AUC</th>
<th>Average AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Y. pestis</em></td>
<td>0.8351</td>
<td>0.8393</td>
<td>0.8310</td>
<td>0.8351</td>
</tr>
<tr>
<td><em>P. ubique</em></td>
<td>0.7355</td>
<td>0.7473</td>
<td>0.7498</td>
<td>0.7442</td>
</tr>
<tr>
<td><em>C. crescentus</em></td>
<td>0.7593</td>
<td>0.7687</td>
<td>0.7585</td>
<td>0.7622</td>
</tr>
<tr>
<td><em>Cyanothece</em></td>
<td>0.7463</td>
<td>0.7469</td>
<td>0.7433</td>
<td>0.7455</td>
</tr>
<tr>
<td><em>Mus musculus</em></td>
<td>0.7473</td>
<td>0.7443</td>
<td>0.7455</td>
<td>0.7457</td>
</tr>
</tbody>
</table>

Figure 2. ROC curves for *Y. pestis*, *P. ubique*, *C. crescentus*, *Cyanothece* and *Mus musculus*.
Chapter 5 – DISCUSSION AND CONCLUSION

A computational technology that improves the efficiency of high-throughput protein identification could lead to several benefits: (1) the identification of true proteotypic peptides not recognized by the classic database searches due to known changes in protein structure from incomplete fragmentation or post-translational modifications (PTM), (2) the reduction of the spectra search space to a subset of computationally predicted proteotypic peptides, and (3) the rapid in silico development of AMT databases for a broad range of organisms.

We have shown that the SVM trained with Y. pestis data accurately predicted proteotypic peptides for phylogenetically diverse prokaryotes, as well as for the mouse. The AUC values reported for these diverse species further encourage the verification and possible improvement of the STEPP classifier. Further development will include research targeting training and testing of SVM models for various protein sample preparation techniques and MS technologies.
REFERENCES


