Applied Bioinformatics (DD2404)

Final Project

Predicting Signal Peptides

Helder Martins, Rendani Mbuvha
Sudhanshu Mittal

Abstract

As part of this project for the course *Applied Bioinformatics* (DD2404) at KTH Royal Institute of Technology we investigate several Machine Learning techniques to build a peptide signal classifier.

In order to utilize the information hidden in these newly discovered protein sequences, it is highly desired to develop computational methods for efficiently identifying their various attributes because the information thus obtained will be very useful for both basic research and drug development. In this project, we build a signal peptide classifier based on the given dataset and further test it on the proteome of two organisms.

In this project, we implement various algorithms to classify which protein sequences start with signal peptide sequence and find that random forest classifier works the best amongst the contemporary classification algorithms with an classification accuracy of 86% on validation dataset including both transmembrane and non-transmembrane sequences. Further, the trained classifier is tested on two BioMart datasets of human and mouse proteomes for signal peptide prediction. In Homo Sapiens(human), proteome sequences 48% were predicted to have signal peptides and in Mus(mouse) sequences 44% were predicted to have signal peptides.

1 Introduction

This project was performed as part of the course *Applied Bioinformatics*(DD2404) at KTH Royal Institute of Technology in Stockholm, Sweden. In this work, we applied machine learning techniques for creating a classifier for peptides signals.

A signal peptide is typically between 15-40 amino acid long and essential for protein secretion. Secreted proteins and a majority of cell-surface possess an N-terminal signal peptide. Secreted and cell-surface proteins are fundamental to inter-cellular communications for multicellular organisms. The extracellular accessibility of these proteins makes them ideal targets for protein therapeutics. In fact, virtually all protein-based therapeutic drugs on the market target these secreted and cell-surface proteins or are secreted proteins themselves.
The importance of signal peptide-containing proteins has motivated the development of several computational methods for predicting signal peptides and determining the signal cleavage sites. Actually, signal peptides have become a crucial tool for pharmaceutical scientists who genetically modify bacteria, plants, and animals to produce effective drugs. The prediction of signal peptides in different secretory proteins is difficult process not only because they are different in sequence components and sequence orders but also in sequence lengths. Another major challenge in the signal peptide prediction is due to the hydrophobic region (h-region) because it can look similar to transmembrane structure.

In this project, we describe and compare various methods to predict whether a protein sequence contains a signal peptide by using machine learning classification algorithms including K-Nearest Neighbor, Support Vector Machines , Decision Tree, Random Forest, Adaptive Boosting and Artificial Neural Networks for transmembrane(TM) and non-transmembrane(non-TM) sequences. In the subsequent sections, we discuss about the structure of signal peptides, related work done in past two decades, about our model designs and experiments. In the results section, we study different behaviour of above mentioned classifiers on TM and non-TM proteins.

1.1 Structure of Signal Peptides

Signal peptides for the secretory pathway generally consist of the following three domains: (i) a positively charged n-region, (ii) a hydrophobic h-region and (iii) an uncharged but polar c-region. The cleavage site for the signal peptide is located in the c-region. However, the degree of signal sequence conservation and length, as well as the cleavage site position, varies significantly between different proteins. The positively charged region (N-region) is 1-5 residues long, a central hydrophobic part (h-region) is 7-15 residues, and a more polar carboxy-terminal domain (c-region) is 3-7 residues long. There is tendency to find small and uncharged residues few positions further from the cleavage site of signal peptide.

1.2 Related Work

During past two decades, variety of predictors have been developed to address this problem. A few recently developed successful techniques have been mentioned below.
PrediSi [7] and SignalP [8] are two popular web-server predictors developed recently for

identifying the signal peptide and its cleavage site. Signal-3L[6] is a 3-layer predictor developed in 2007 for predicting signal peptide sequences and their cleavage sites in human, plant, animal, eukaryotic, Gram-positive, and Gram-negative protein sequences, respectively. Another recent developed method [3] is a 2-layer predictor: the 1st-layer prediction engine is to identify a query protein as secretory or non-secretory; if it is secretory, the process will be automatically continued with the 2nd-layer prediction engine to further identify the cleavage site of its signal peptide.
According to the recent survey report [2], Signal-CF performed the best in predicting the long signal peptides, among the following eight web-server predictors: SignalP-NN [8], SignalP-HMM [8], Sig-nalP-NN or SignalP-HMM [8], Phobius [9], PrediSi [7], Signal-CF [3], Signal-3L [6], and Philius [10].

2 Experiment Setup

2.1 Data-set

The dataset consists of pre-selected protein sequences evaluating to positive and negative for the signal peptide presence. The samples are further split based on the properties of their membrane: transmembrane and non-transmembrane proteins. There are 2362 samples in the non-transmembrane subset, of which 54% contain a signal peptide, and 292 samples in the transmembrane subset, where only 15% evaluates to positive for the presence of a signal peptide. The transmembrane subset of the dataset is highly unbiased, therefore it is harder to learn an unbiased classifier over such data.

2.2 Sequence Logs

In order to visualise what a typical signal peptide would look like we make use of sequence logos. Sequence logos are a representation of relative consensus and diversity in a dataset. In our cases, the logo depicts the relative frequencies of different amino acids at certain positions in the sequences. The logos were created using UC Berkeley's Weblogo [4] online tool.

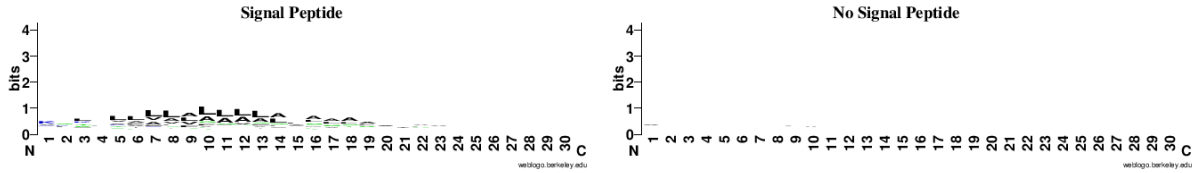

Figure 1: Sequence logos for sequences with a signal peptide(left) and sequences without signal peptides(right). For ease of visualisation the first amino acid which is heavily weighted on Methionine is removed and sequences were limited to the first 30 amino acids. Notice: The sequence logos are very sparse in the sequences without signal peptides.

From figure 1 we can see that sequences with signal peptides tend to have a high densities of amino acids Lysine (L) and Alanine (A) in postions 5 to 20. Sequences without peptides tend not have any dominant amino acids.

2.3 Design Choices

For our signal peptide classifier we cross-validated several different classification algorithms to evaluate which would result in a good accuracy for this specific task. The classifiers used where: K-Nearest Neighbours, Support Vector Machines (with linear and radial-basis kernels), Decision Tree, Random Forests, AdaBoost and Multi-layer perceptron. For each one of the classifiers, a model was trained using 75% of the original data, while the remaining 25% where used for testing.

One important part of the training process is deciding how our data will be represented, feeding long string sequences to a model without any consideration for its structure often results in poor performance. In this project, the amino acid sequence of a protein is first fed to a K-gram vectorizer which stores the counts of how many times a subsequence of size K is found. This transformation is based on the hypothesis that the signal peptide presence is influenced by the presence of a subsequence in the protein. This transformation, while sensible, has the undesirable effect of losing the exact position of the subsequence, since we are only interested in how many times it occurs. Knowing that the signal peptide is commonly found in the beginning of the sequence, we can filter the first 60 aminoacids as done by [8] as to improve the accuracy of our classifiers and mitigate the loss of information caused by the transformation.

Training the models directly with K-grams is unfeasible since the large amount of features of each sample would require a prohibitively large number of training data as to obtain a decent accuracy for the classifiers, which is known in Machine Learning as the "Curse of Dimensionality". To avoid this problem, feature reduction techniques should be used as to keep only the latent attributes that pinpoint the presence of the signal peptide. In this project, two technique were used: Feature Hashing [1] followed by a Truncated Singular-value Decomposition (SVD) [5]. Feature Hashing maps the K-grams sub-sequence to a value using a hashing function while allowing some of these values to collide (thus reducing the dimensionality), while SVD further reduces the "Samples $\times$ Features" matrix by decomposing it and using the vectors corresponding to the largest singular values.

Another design choice that yielded a significant improvement was boosting the H-region of the positive samples. It is known that the hydrophobic region has the largest statistical influence on the presence of the signal peptide, so our approach increases the weight of all amino acids in this region by replicating them a configurable amount of times. In practice this means that this part of the sequence will be counted more times by the K-gram vectorizer, thus implying that the subsequence was present more times that it actually was.

3 Results

3.1 Performance on Transmembrane vs Non-Transmembrane Proteins

In Figure 2, the plots derived from the models trained on the non-transmembrane dataset are given. With an empirical assessment of the arguments used on the data transformation it was verified that the best size of the K-grams is ranging from 3 to 6, and the H-region amino acids are weighted at 3 times the ones outside the region. The Support Vector Machine classifier with a radial-basis kernel (rightmost plot) and the Multilayer

Perceptron are the classifiers who performed best while also maintaining a good balance between the prediction accuracy on both positive and negative samples.

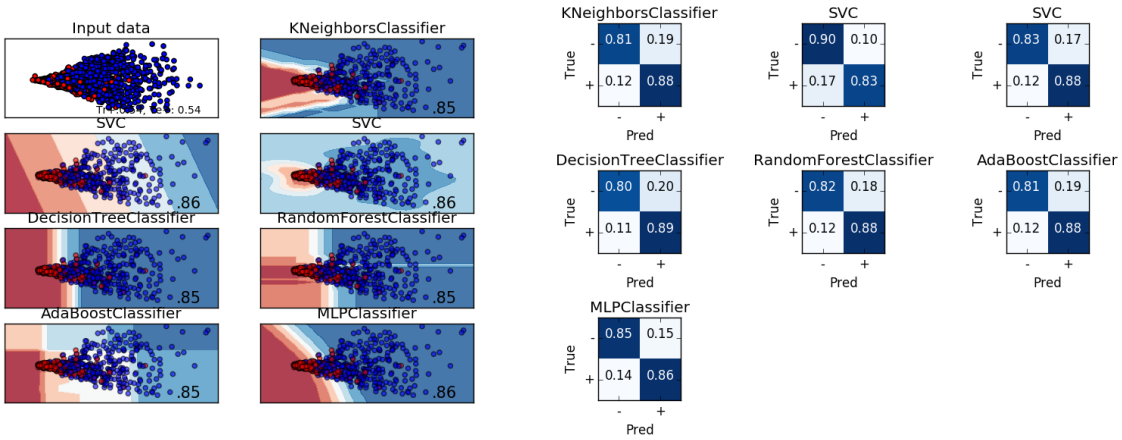

Figure 2: Classifiers and confusion matrices for the non-transmembrane dataset. The input data plot are the samples used for training. The classifiers are plotted with the test samples together with its decision boundary and their accuracy on a test set. Positive samples are colored blue. Note: The accuracy of classification is mentioned at the right bottom of the classifier plot.

In Figure 3, the plots derived from the models trained on the transmembrane dataset are given. With an empirical assessment of the arguments used on the data transformation it was verified that the best size of the K-grams is ranging from 6 to 10, and the H-region amino acids are weighted at 3 times the ones outside the region. The Support Vector Machine classifier with a radial-basis kernel (rightmost plot) and the Random Forest are the classifiers who performed best with an identical accuracy, while also maintaining a good balance between the prediction accuracy on both positive and negative samples.

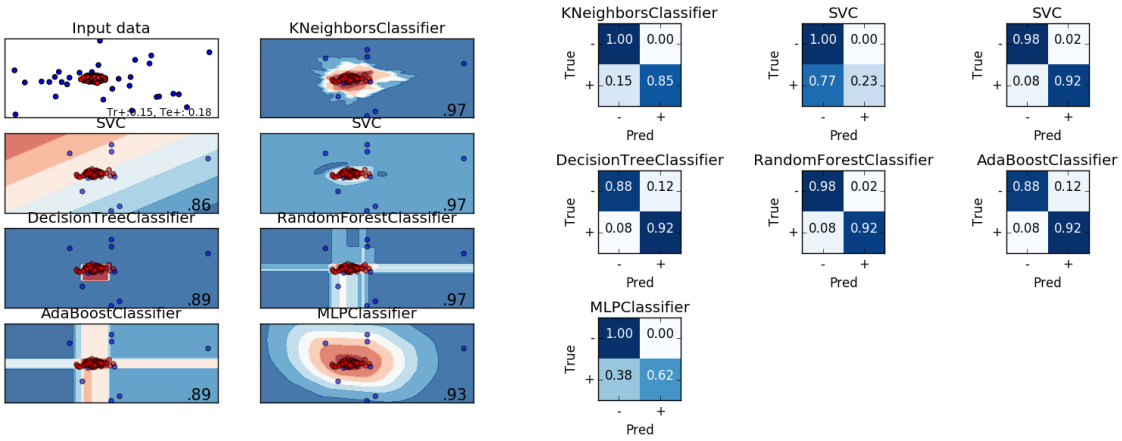

Figure 3: Classifiers and confusion matrices for the transmembrane dataset. The input data plot are the samples used for training. The classifiers are plotted with the test samples together with its decision boundary and their accuracy on a test set. Positive samples are colored blue. Note: The accuracy of classification is mentioned at the right bottom of each classifier plot.

3.2 Signal Peptide Detection in Human and Mouse Proteomes

We further used the classifiers trained in the previous section to detect signal peptides in proteomes of Homo sapiens (Human) and Mus musculus (Mouse). The datasets in fasta format were obtained from Ensemble's BioMart service's ftp site. The datasets contained 102915 protein sequences for Human and 61440 for Mouse sequences. Each dataset was transformed into same K-gram vectorizer as was used in training. The resulting features are then scored through each classifier to produce following prediction results. Tables 1 and 2 show results for Homo Sapiens and Mus musculus proteomes respectively.

Table 1: Classification Results on Homo sapiens proteome

Classifier	Predicted Peptides Signal	Predicted% of Signal Peptides
KNN	46124	45%
SVC -Linear	30143	29%
SVC -RBF	41710	41%
Decision Tree	45830	45%
Random Forest	48916	48%
Adaboost	49505	48%
MLP	38492	37%

Table 2: Classification Results on Mus musculus proteome

Classifier	Predicted Peptides Signal	Predicted% of Signal Peptides
KNN	26203	43%
SVC -Linear	17043	28%
SVC -RBF	23378	38%
Decision Tree	25121	41%
Random Forest	27139	44%
Adaboost	25517	42%
MLP	21715	35%

t

4 Discussion and Concluding Remarks

Testing of our trained signal peptide classifiers on unseen data results in relatively high accuracies of atleast 80%. Noteably the Random Forest, Radial basis function based Support Vector Machines and the Multilayer Perceptron are the top performers. However, the multilayer perceptron and Linear SVM tend to underperform in the transmembrane dataset for positive predictions. This decrease in sensitivity can also be attributed to significantly reduced size of the trans-membrane dataset relative to the non-transmembrane. The absence of a cleavage site for transmembrane domain might also have contributed to this decline. The random forest maintains the most consistent positive performance in classification accuracy.

There is significant change in the behaviour of classifier in non-transmembrane and transmembrane datasets. In the transmebrane dataset, the sensitivity decreases while the specificity decreases. This is due to the inherent bias in the data where the prevalence of signal peptides in non-trans-membrane is 54% but only 15% in transmembrane.

When applied to the Human and Mouse proteomes the Random Forest classifier predicts that 48% and 44% of the proteins contain signal peptides in each species respectively.

There are multiple possible further improvements to this work. First variations in feature extraction can be tested from simple counts of amino acids to more complex higher order n-grams that were not presented here. In addition to this the possibility of using additional principal components for training can also be explored. Secondly, other models that can handle the spatial sequencing aspect of that data, can also be explored such as Hidden Markov Models and Recurrent Neural Networks to obtain better classification performance.

References

[1] Cornelia Caragea, Adrian Silvescu, and Prasenjit Mitra. Protein sequence classification using feature hashing. *Proteome science*, 10(1):1, 2012.

[2] K. Chou and H. Shen. Review : Recent advances in developing web-servers for predicting protein attributes. *Natural Science*, 1(63-92), 2009.

[3] K.C. Chou and H.B. Shen. Signal-cf: a sub-site-coupled and window-fusing approach for predicting signal peptides. *Biochem Biophys Res Comm*, 357(633-640), 2007.

[4] Gavin E. Crooks, Gary Hon, John-Marc M. Chandonia, and Steven E. Brenner. WebLogo: a sequence logo generator. *Genome research*, 14(6):1188–1190, June 2004.

[5] F Fogolari, S Tessari, and H Molinari. Singular value decomposition analysis of protein sequence alignment score data. *Proteins: Structure, Function, and Bioinformatics*, 46(2):161–170, 2002.

[6] K.C. Chou H.B. Shen. Signal-3l: A 3-layer approach for predicting signal peptide, biochem. *Biophys. Res. Commun*, 363(297303), 2007.

[7] Scheer M Mnch R Jahn D. Hiller K, Grote A. Predisi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Research*, 32(Web Server issue):W375–W379; doi:10.1093/nar/gkh378., 2004.

[8] G. von Heijne S. Brunak J.D. Bendtsen, H. Nielsen. Improved prediction of signal peptides: Signalp 3.0. *J. Mol. Biol.*, 340(783795.):W375–W379; doi:10.1093/nar/gkh378., 2004.

[9] Krogh A. Kall, L. and E.L. Sonnhammer. Ad-vantages of combined transmembrane topology and sig-nal peptide prediction–the phobius web server. *Nucleic Acids Res*, 35(W429-43), 2007.

[10] Kall L. Riffle M.E. Bilmes J.A. Reynolds, S.M. and W.S. Noble. Transmembrane topology and signal peptide prediction using dynamic bayesian networks. *PLoS Comput Biol*, 4(e1000213), 2008.