# Applied Bioinformatics (DD2404)

Project Report

# Predicting Signal Peptides

Prashant Kumar

## Abstract

The project *Predicting Signal Peptides* was done as a part of the course Applied Bioinformatics (DD2404) at KTH Royal Institute of Technology where I have experimented with various available machine learning models at SciKit learn to build a classifier to predict signal peptides.

To signify the essence of the project, it is important to utilize the implicit information present in the given protein sequence which could be utilized for both basic research and drug enhancement. This demands for a computational method that could identify important features, attributes in the protein sequence. As a part of this, I build a signal peptide classifier and train It on the training dataset available on the course webpage and test it on the proteome of two organisms. To dive into this further, I implement different ML algorithms to classify which protein sequences start with signal peptide sequence. As a result of it, I conclude that *Random Forest* classifier provides better classification accuracy of 86% on validation dataset including both Transmembrane and Non-Transmembrane sequences. Following that, the trained classification model was tested over two BioMart datasets of *human* and *mouse* proteomes for signal peptide prediction. In Homo Sapiens(human), 49% of proteome sequences were predicted to have signal peptides and in Mus(mouse) sequences 45% were predicted to have signal peptides.

## 1 Introduction

Briefing into the project, Signal peptides are extra peptide extension that contains 15–40 amino acids added to the N-terminus of the protein. They work by assisting the transport functionality within the cell to bring it to its specific destination which protein is delivered within the cell. Secreted and cell-surface proteins are important to intercellular 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. It is removed while the protein is translocated across the endoplasmic reticulum membrane.

The value of signal peptide-containing proteins has initiated the development of various computational methods for predicting signal peptides and determining the signal cleavage sites. The prediction of signal peptides in different secretory proteins is an arduous task process because they are dissimilar in sequence components, sequence orders in addition to the sequence lengths. Another major problem to overcome in the signal peptide prediction is that the hydrophobic region (h-region) can look similar to the transmembrane structure.

In this project, I perform a classification analysis of different suitable ML classifier algorithms to predict whether a protein sequence contains a signal peptide. The algorithms considered here are K-Nearest Neighbour, Support Vector Machines, Decision Tree, Random Forest, AdaBoost and Artificial Neural Networks for transmembrane(TM) and non- transmembrane(non-TM) sequences. In the following sections, we elaborate about the structure of signal peptides, related work done in the similar field, explanation about the models implemented and experiments conducted over the datasets. In the results section, an analytical study of different behaviour of above mentioned classifiers on TM and non-TM proteins is presented.

## 1.1 Structure of Signal Peptides

Signal peptides for the secretory pathway usually consist of the following three domains: (i) a *positively* charged n-region which is 1-5 residues long, (ii) a central *hydrophobic* h-region which is 7-15 residues and (iii) an *uncharged* but polar c-region which is 3-7 residues long. 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, differs significantly between various proteins. There is tendency to find small and uncharged residues few positions further from the cleavage site of signal peptide.

## 1.2 Related Work

A lot of work has been done in this domain where variety of predictors have been developed to address this problem. A few recently implemented techniques where referred for this work which has been subsequently mentioned in the reference section below such as PrediSi [7] and SignalP [8]. 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. 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], SignalP-NN or SignalP-HMM [8], Phobius [9], PrediSi [7], Signal-CF [3], Signal-3L [6], and Philius [10]. Another 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.

## 2 Experiment Setup

### 2.1 Dataset

The dataset consists of preselected protein sequences evaluating to positive and negative for the signal peptide presence. The data 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 given such dataset.

### 2.2 Sequence Logs

Sequence logos are a representation of relative consensus and diversity in a dataset. In the given case, the logo depicts the relative frequencies of different amino acids at certain positions in the sequences. The logos were implemented using UC Berkeley's Weblogo [4] online tool.

Figure 1: Sequence logos for sequences with a signal peptide(right) and sequences without signal peptides(left). 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.

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 positions 5 to 20. Sequences without peptides tend to not have any dominant amino acids.

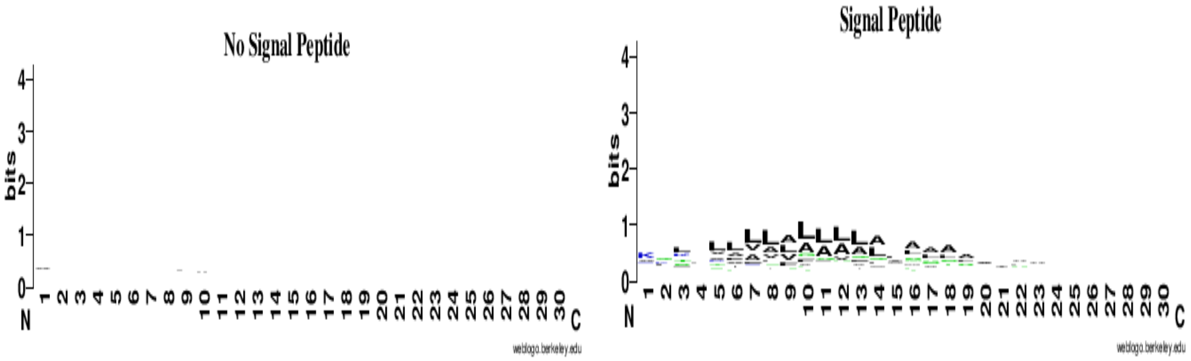


*Figure 1: Sequence logos*

### 2.3 Design Experiments

For the corresponding classifier, I cross-validated several classification algorithms like K-Nearest Neighbours, Support Vector Machines (with linear and radial-basis kernels), Decision Tree, Random Forests, AdaBoost and Multi-layer perceptron to find out the best classifying algorithm. For each one of the classifiers, a model was split into train and test dataset in a 70/30 ratio.

A critical aspect of training was the method of representing the data to the model, as feeding long string sequences to a model without any consideration for its structure often resulted in undesired 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 driven by the hypothesis that the signal peptide presence is influenced by the presence of a subsequence in the protein. Although, it has an undesirable effect of losing the exact position of the subsequence, since we are only interested in count of its occurrence. Knowing that the signal peptide is commonly found in the beginning of the sequence, we can filter the first 60 amino acids as done by [8] as to improve the accuracy of our classifiers and alleviate the loss of information caused by the transformation.

Training the models directly with K-grams is undesirable as it could suffer from the “Curse of Dimensionality”. To avoid this problem, dimensionality reduction techniques such as PCA, Feature Hashing, ICA etc. should be used as to keep only the latent attributes that pinpoint the presence of the signal peptide. In this project, two techniques 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 therefore resulting in dimensionality reductions, while SVD further reduces the

“Samples × Features” matrix by decomposing it and using the vectors corresponding to the largest singular values.

Another design choice that yielded a considerable 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 the approach increases the weight of all amino acids in this region by replicating them a configurable amount of times. In simple terms, this means that this part of the sequence will be counted more times by the K-gram vectorizer.

## 3 Results

### 3.1 Classifier Performance on Transmembrane vs Non-Transmembrane Proteins

In Figure 2, the plots represent the models trained on the non-transmembrane dataset are shown. With an empirical assessment of the it was found 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 and MLP are the classifiers who performed best while also maintaining a good balance between the prediction accuracy on both positive and negative samples.

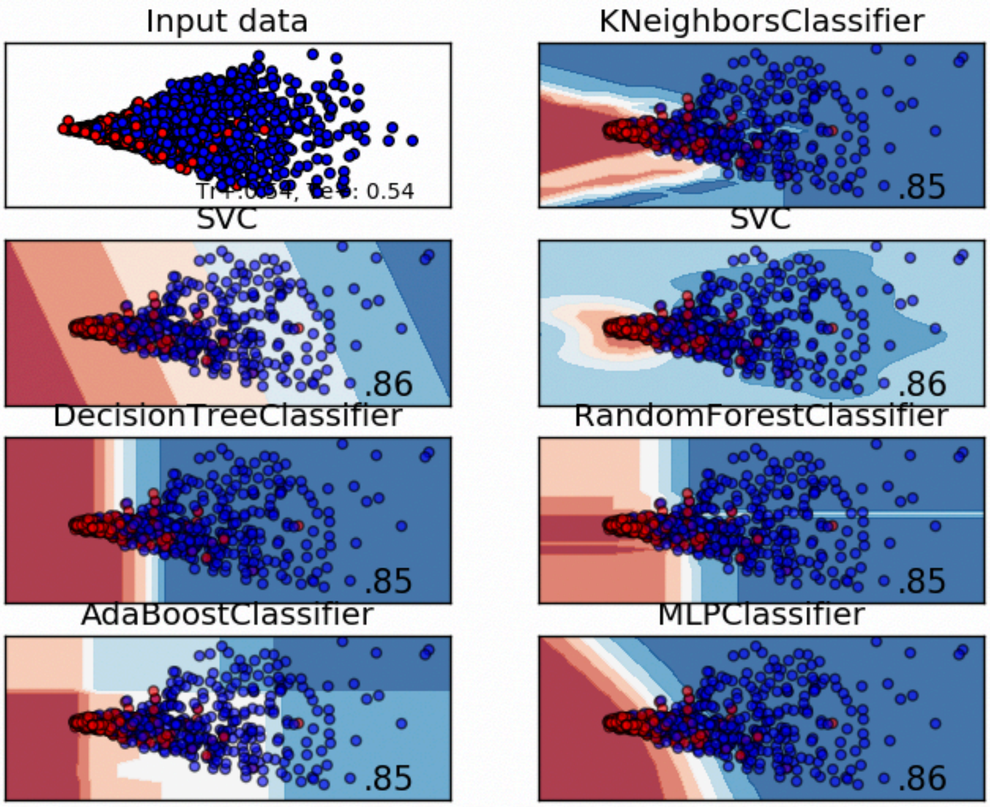


*Figure 2: Models trained on the non-transmembrane dataset*

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.

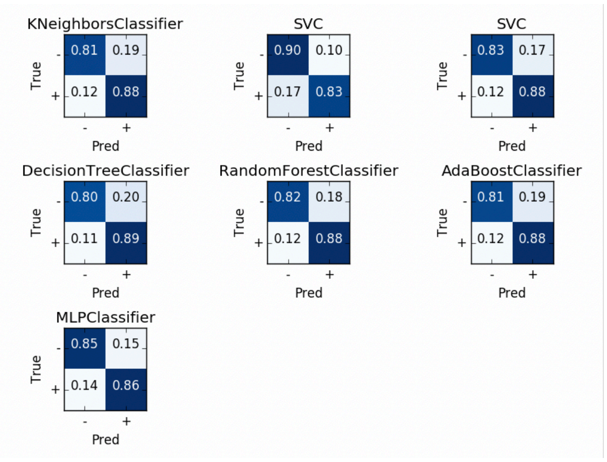
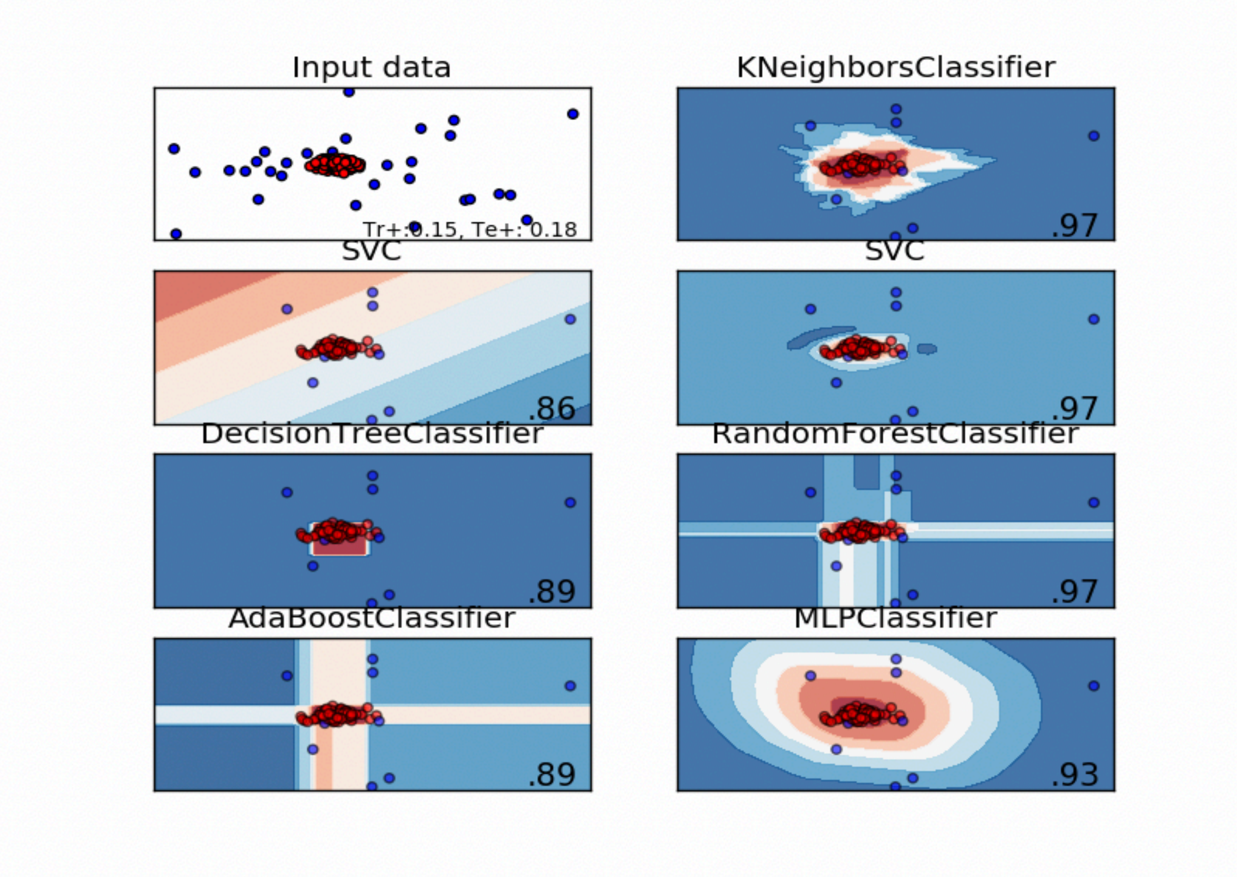
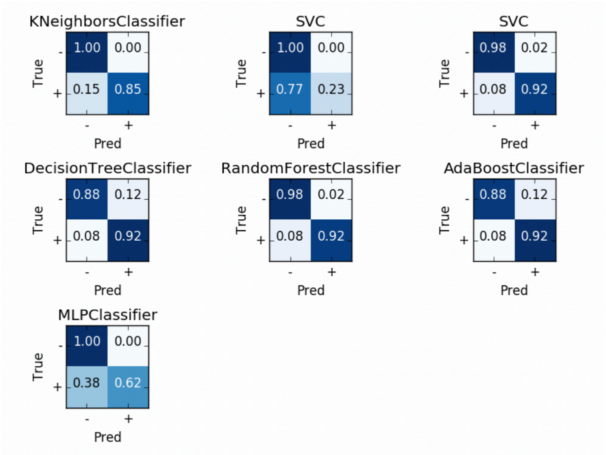


*Figure 3: Classifiers and confusion matrices for the non-transmembrane dataset*

The Support Vector Machine classifier with a radial-basis kernel 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.



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.



### 3.2 Testing Model on Test Set (Signal Peptide Detection in Human and Mouse Proteomes)

The trained model was then tested over the Human and Mouse proteomes dataset. The datasets are in fasta format that 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 |
|---|---|---|
| SVC-RBF | 41710 | 42 |
| KNN | 46124 | 46 |
| SVC Linear | 30143 | 29 |
| Decision Tree | 45830 | 46 |
| Random Forest | 48916 | 49 |
| MLP | 38492 | 49 |
| AdaBoost | 49505 | 37 |

Table 2: Classification Results on Mus musculus proteome

| Classifier | Predicted Peptides Signal | Predicted % of Signal Peptides |
|---|---|---|
| SVC-RBF | 23378 | 39 |
| KNN | 26203 | 44 |
| SVC Linear | 17043 | 27 |
| Decision Tree | 25121 | 42 |
| Random Forest | 27139 | 45 |
| MLP | 21715 | 34 |
| AdaBoost | 25517 | 41 |

## 4 Discussion and Conclusion

The performance of the classifier over the test data set gave an accuracy 80% with Random forest, RBF based SVM and MLP were the best giving classification results. However, the MLP and Linear SVM tend to underfit the transmembrane dataset for positive predictions which could be attributed to significantly diminished size of the transmembrane dataset relative to the non-transmembrane. In addition to that, 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 which 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 49% and 45% of the proteins contain signal peptides in each species respectively

## 5. Future Work:

Various variations in feature extraction can be experimented by starting with simple counts of amino acids to a more complex higher order n-grams that wasn't presented here. Also, there is a scope of using principal components to find out the direction of highest variance in the data. With advancement of ML, this problem could have also been tackled with HMM or even Recurrent Neural Network which could have given a much better accuracy.

### 6. References

[1] Cornelia Caragea, Adrian Silvescu, and Prasenjit Mitra. Protein sequence classification using feature hashing. Proteome science.

[2] K. Chou and H. Shen. Review : Recent advances in developing web-servers for predicting protein attributes. Natural Science.

[3] K.C. Chou and H.B. Shen. Signalcf: a sub-site-coupled and window-fusing approach for predicting signal peptides. Biochem Biophys Res Comm.

[4] Gavin E. Crooks, Gary Hon, John-Marc M. Chandonia, and Steven E. Brenner. WebLogo: a sequence logo generator. Genome research.

[5] F Fogolari, S Tessari, and H Molinari. Singular value decomposition analysis of protein sequence alignment score data. Proteins: Structure, Function, and Bioinformatics

[6] K.C. Chou H.B. Shen. Signal-3l: A 3-layer approach for predicting signal peptide, biochem. Biophys.

[7] Scheer M Mnch R Jahn D. Hiller K, Grote A. Predisi: prediction of signal peptides and their cleavage positions.

[8] G. von Heijne S. Brunak J.D. Bendtsen, H. Nielsen. Improved prediction of signal peptides.

[9] Krogh A. Kall, L. and E.L. Sonnhammer. Ad-vantages of combined transmembrane topology and

signal peptide prediction—the phobius web server.

[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.