Project Report

Comparing vonHeijne and SVM in predicting protein signal peptides

Taccaliti Edoardo

Abstract

Motivation: The project aims to compare two deterministic methods: Gunnar Von Heijne model and SVM for predicting secretary signal peptides in proteins which is of primary importance in localizing protein compartment and functional characterization.

Results: After a comparative analysis on benchmark set of the two developed methods for signal peptides detection and despite some misclassification errors, SVM is the best performing method having an higher MCC, precision and F1 score.

Supplementary information: report's attachments – "Supplementary materials" folder

1 Introduction

Signal peptides (SP) are short peptides located in the N-terminal of proteins, carrying information for protein secretion. They are ubiquitous to all prokaryotes and eukaryotes. SPs have been of special interest in several scientific and industrial fields, including recombinant protein production, disease diagnosis, immunization, and laboratory techniques. The modular architecture of SP accounts for three distinguished regions: A positively charged N-region, the central hydrophobic region (H-region) and a polar uncharged C-region hosting the cleavage site (A-X-A consensus motif for the signal peptidases enzymes (Von Heijne, 1986) .The cell-machinery exploits its peculiar structure in redirecting proteins into cell's compartment making the SP in-silico recognition a key step for the characterization of protein function and subcellular localization.The interest in the automated identification is related to the huge amount of unprocessed data available and to the aim of complementing experimental results in genome scale analysis, but also to the industrial need to find more effective vehicles to produce proteins in recombinant systems, or to vaccine development (Nielsen et al., 1997). Moreover, the knowledge in protein localization allows thee identification of potential protein interactors. One major challenge in discriminating true signal peptides is represented by hydrophobic regions, since the cleavage-site pattern alone is not sufficient to distinguish between signal peptides and these types of sequences, such as Transmembrane Alpha-helices and transit peptides for subcellular compartments like Chloroplast, mitochondrial and peroxisomal. This is an underlying problem in scanning sequences for the presence of signal peptides, which will yield a lot of false positive predictions in the N-terminal domain (Petersen et al., 2011). SignalIP-6 is the current state of the art method in predicting, given its ability to predict both proteins used to make the model and unknown proteins from metagenomic data. Differently from the predictors already present in the panorama, SP-6 does not need to know the origin of the data. (Teufel et al., 2022). The first developed method for the in-silico prediction of signal peptides was introduced by Gunnar von Heijne in 1983 (VON HEIJNE, 1983) , and it was based on a reduced-alphabet weight matrix used to model the cleavage site region. Nowadays, different methods for SP prediction are available in the scientific panorama, ranging from machine learning based methods, to deep learning.

Herein, we compare two different methods in predicting Signal Peptides. An updated successive version of the vonHeijne weight matrix-based approach (Von Heijne, 1986) , and SVM or Support vector machines, a machine learning based method (Cai et al., 2003). These two methods have been trained and tested on the same benchmark dataset derived from SignalP-5.0 to evaluate their performances in discriminating signal peptide sequences. From the results, despite both methods were making some misclassification errors, the SVM, is the best performing method overall.

2 Methods

2.1 Datasets:

The datasets used for this project were already provided from courtesy of Professor Savojardo.The training dataset was derived from SignalP-5.0 datasets (Alamagro Armenteros et al,2019), extracted from uniport Knowledgebase release 0f 2018/04. Which comprises 1723 eukaryotic sequences with following requirements: Only reviewed (SwissProt); proteins sequences longer than 30aa; only signal peptides with experimental evidence code (ECO: 0000269) for the cleavage site included. Of the 1723 sequences, 258 are positive examples having an N-terminal secretory signal peptide. 1465 sequences are negative examples, proteins annotated with a subcellular location, in addition with a randomly selected subset of the original SignalP-5.0 training dataset. The benchmark set was the same dataset adopted by SignalP-5.0 and it contains 7456 eukaryotic sequences of which 209 are positive examples and the remaining 7247 are negative examples.

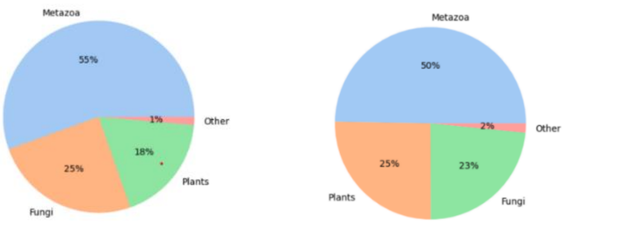
2.2 Dataset pre-processing:

For our purpose, only the first 50 N-terminal residues of all sequences were kept. Data was split according to pairwise sequence similarity (>30%), into 5 equally sized different subsets in order to perform cross-validation. In this way, redundancy between subsets was avoided and hence a possible bias between training and testing sequences.

2.3 Dataset statistics:

Both training and testing datasets were investigated by plotting SP lengths distribution [Fig.1.1 and Fig1.2. supplementary materials] and Taxa distributions. [Fig.2.1 and Fig2.2 supplementary materials]. Moreover, the amino acid compositions of both datasets with respect to the average amino acid composition of SwissProt-database [Fig.3 and Fig.4. supplementary materials] Signal peptide distribution among kingdoms was investigated via pie chart, highlighting a prevalence of kingdom metazoan (50%), plants (around 25%) and fungi (20%) Fig.5. In addition, sequence logo profiles(Crooks et al., 2004) were computed for both datasets [Fig.6.1 and Fig.6.2 supplementary materials], showing similarity between the two datasets.

Fig.5. Kingdom distribution:



(On the left) kingdom distribution training and benchmark (on the right).

2.2 vonHeijne method:

Is a simple discriminative method for motif discovery and recognition. It uses a weighted matrix, called Positional-Specific Weight Matrices or (PSWM) a statistical method for modeling the region around the cleavage site where rows are equal to the number of different amino acids and the columns equal to the length of the motif. It works by converting the observed number of each of the residues in each position of the aligned signals into a measure of probability of finding that residue in each position of the sample.

From a dataset of proteins endowed with SPs only, a PSPM was computed from the region surrounding the cleavage-sites, covering in total 15 residues: From -13 residues downstream respect to the cleavage site, to +2 residues upstream, therefore modelling the most hydrophobic portion of the signal peptide. Once the matrix was finalized with all the occurrence scores, each position was divided by N+20, with N equal the length of signal peptide. Then divided by the frequency of SwissProt-Background distribution. Finally, the natural logarithm was computed. Given the fact that not all the residues are present in each position and to avoid zero probabilities, pseudocounts were set a priory, assuming each residue was spotted at least once in all positions.

The following formula was used to compute PSPM:

$$M_{k,j} = \frac{1}{N+20}\left(1 + \sum_{i=1}^{N} I\left(s_{i,j} = k\right)\right)$$

Where:

- $s_{i,j}$ is the observed residue of aligned sequence i at position j.
- k is the residue corresponding to the k-th row in the matrix.
- $I\left(s_{i,j} = k\right)$ is the indicator function (if condition is met 1, otherwise is 0)

From a PSPM, a PSWM was computed starting from the probability of occurrence of one residue over the background distribution (SwissProt).

$$W_{k,j} = \log \frac{M_{k,j}}{b_k}$$

The computed PSWM was then used to predict the presence or absence of Signal peptide on a group of testing sequences. Therefore, a window of 15 residues long was implemented and slide one position at the time. For each sliding loop, the score given by the formula below was calculated, the higher the score, higher the probability that the sequence was a Signal peptide.

$$score\ (X\ |W) = \sum_{i=1}^{L} Wx_i, i$$

To assign a positive or negative class to each sequence, a threshold was computed using a 5-times cross validation procedure. 4 folds were used to estimate the threshold, leaving the last one (validation set) for prediction. During each run, an optimal threshold was calculated using precision and recall values. In the end, the mean between the predicted threshold was found to be **8.206**. This threshold along with a PSWM designed using positive sequences of the training set will be used to test unseen sequences of the benchmark set.

2.3 Support Vector Machines

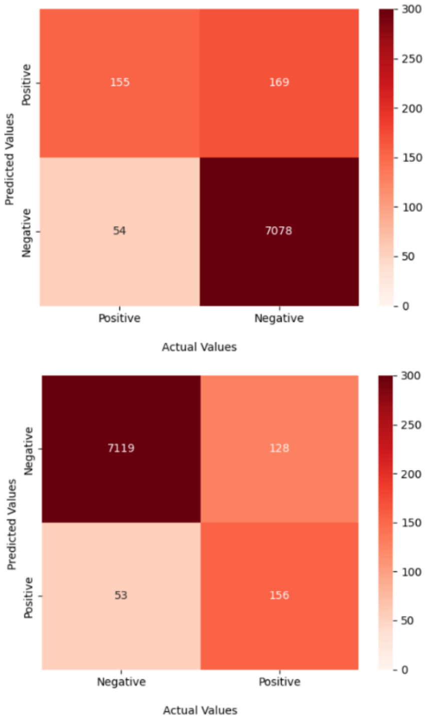
Support Vector Machines or SVM are a supervised method for classification and regression. The main motivation behind SVM is to linearly separate classes in the training set by a surface that maximizes the margin between them. To achieve this, support vectors must be identified, which are the closest points to the hyperplane or decision boundary and where the margins maximizing the separation are draw. SVM are simple and powerful tools for linearly classifying points. Radial basis function or (rbf) are a type of SVM extremely fitted for this purpose. It highlights similar points, those having the least distance by encircling them into an island. In the case datapoints could not be properly linearly separable, a soft margin SVM introduce a slack variable to allow datapoints misclassification and at the same time maximizing the margin separator plane. For points where the separation in a 2-Dimensional space is not feasible, the original feature space could be mapped to a higher dimensional space, the feature space, where the datapoints becomes linearly separable, this is known as 'kernel trick'.
The kernel function is a function equivalent to an inner product in the feature space that allow to map data to the high-dimensional space without computing the feature space explicitly. For our purposes, we choose a non-linearly separable SVM implemented in scikit-learn library. (Pedregosa FABIANPEDREGOSA et al., 2011)

Firstly, psycho-chemical characteristics of the Signal peptide that could be discriminative for the SVM were selected. By exploiting the median length of SPs obtained from length distribution, input sequences in both training and testing were represented as a datapoint in the feature space by 20-dimensional vectors, where each dimension corresponds to the relative frequency of the amino acid itself. The length of the encoded sequences *k* is an hyperparameter that was optimized through a 5-times cross-validation procedure along with other two hyperparameters. It was introduced to deal with the problem of different signal peptides length. From the signal peptide distribution of the training set, it was found that 21 was the mean SP length and 20±5 was the average SP length. Thus, expecting most of the signal peptide having a K parameter ranging between 17 and 24. The penalty parameter C of the error term, which is a trade-off factor between margin maximization and misclassification errors. The kernel coefficient for the RBF kernel or (gamma parameter), which was tested on several values comprising 'scale'. These hyperparameters were also tested on a cross-validation run, the combination that gave the best result in terms of MCC during the grid-search, were the selected ones for benchmark testing.

2.4 Scoring metrices

To evaluate the performances, both methods were tested on the benchmark set. There were 5 scoring measures obtained from the confusion matrices and used for evaluation purposes, which are: MCC, precision, recall, F1 score and Accuracy (Fig.3).

Fig.3. Confusion matrices vonHeijne and SVM



Figures showing parameters derived during 5 times cross-validation. for vonHeijne, (above) and SVM (below).

2.4.1 Matthews Correlation Coefficient

Is the correlation coefficient between observed and predicted binary classification and it measures the difference between predicted and actual values and it is used when negative and positive cases are of equal importance. It summarizes classifier's performances between -1 and +1 (respectively for a poor model and a good model). It is robust to class imbalances.

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

2.4.2 Precision

Precision or positive predicted value is the proportion of correctly classified positive examples. Or is the ability of the model to not label as positive negative examples.

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

2.4.3 Recall

Is the ability of the model to find all the positive examples. It is calculated as

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

2.4.4 F1 score

F1 score is the harmonic mean between Precision and recall, it provides a score ranging between 0 (for bad performances) and 1 (which represent a model able to classify examples in the correct class).

$$\mathrm{F}_1 = 2 \times \frac{\mathrm{PPV} \times \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

2.4.5 Accuracy

Accuracy is the proportion of correct predictions (true positive and true negatives) in all evaluated cases.

$$\mathrm{Accuracy} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}}$$

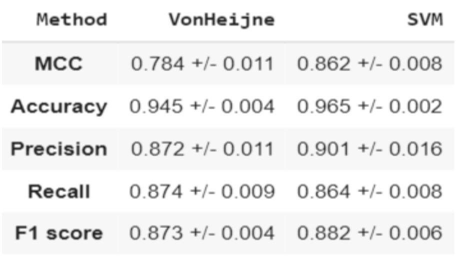
3 Results

3.1 Cross-validation results

To optimize both methods, a cross validation procedure was adopted. In the case of the vonHeijne, CV was used to extract the best threshold score for the prediction against the blind set.

Regarding the SVM, the CV was used for selecting the best hyperparameters. To make a preliminary comparison, in each fold, five evaluation parameters and the standard errors were computed. (Table 1)

Table 1: average performance metrics during Cross-validation.

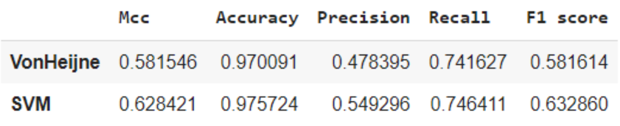


Method	VonHeijne	SVM
MCC	0.784 +/- 0.011	0.862 +/- 0.008
Accuracy	0.945 +/- 0.004	0.965 +/- 0.002
Precision	0.872 +/- 0.011	0.901 +/- 0.016
Recall	0.874 +/- 0.009	0.864 +/- 0.008
F1 score	0.873 +/- 0.004	0.882 +/- 0.006

Table showing the average performance metrics obtained during CV for both methods.

Afterwards, methods were tested over the same benchmark set. In the case of vonHeijne the best threshold was found to be 8.20, whereas for SVM, the best hyper-parameters combination was a window size of 19, a C value of 3 and 'scale' as gamma parameter. Model's performances were calculated over the benchmark. (Table.2)

Table.2: Benchmark performance metrics



	Mcc	Accuracy	Precision	Recall	F1 score
VonHeijne	0.581546	0.970091	0.478395	0.741627	0.581614
SVM	0.628421	0.975724	0.549296	0.746411	0.632860

Table performance metrics benchmark clearly showing MCC, precision and F1 score are higher in SVM.

3.2 False Positives results

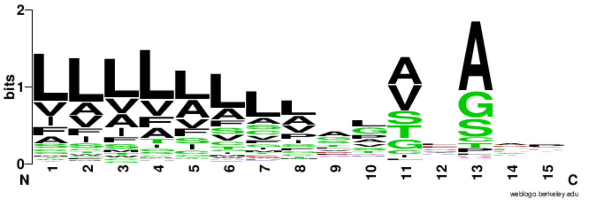
A potential reason behind the generation of false positives may come from those proteins having a Transmembrane alpha-helix region. This is due to the fact that, both SP and Transmembrane regions share a high number of hydrophobic residues in their first 50 residues in the N- terminal. In addition, transit peptides, which are short hydrophobic signals located at the protein N-terminal domain, and regulating export of proteins towards cytoplasmatic organelles, hence: mitochondria, chloroplast, and peroxisomes. Therefore, sequences containing transit peptides are also a possible source of false positives.

To investigate the effect of these proteins on the overall number of negative examples, False positive rates, representing the ratio between false positives and all the negative examples were calculated [Table.3 supplementary materials]. This number which should be the lowest possible, was found to be 0.0233 and 0.0176 respectively for the vonHeijne and SVM algorithms. Moreover, False positive rates specific for Transmembrane alpha-helix were calculated, 0.297 for vonHeijne and 0.3040 for SVM and for sequences containing transit peptides 0.0351 for vonHeijne and 0.0294 for SVM. In addition, false positive rates were computed also for peroxisomes, mitochondria (0.033 and 0.044) and chloroplast (0.040 and 0.0128). The specific localization of the transmembrane alpha-helices and transit peptides at the first 50 N-terminal residues, fostered our model in over-predicting signal peptides. In fact, the false positive results shows that TM domains are the primary source of misclassification, thus, special care should be kept in including these kinds of sequences in signal peptide prediction analysis.

3.3 False Negative results

The cause for the presence of False Negatives should be imputed to the specific design of the two methods. Regarding vonHeijne method, the discrimination principle exploits the 'cleavage-site context', therefore a careful analysis of the sequence logo profiles of true positives, false negatives via comparison with whole training sequences (Fig.7) was carried out to check if the cause behind getting false negatives could be due to the different composition in the context of the cleavage-site.

Fig.7. Sequence logo profile training:

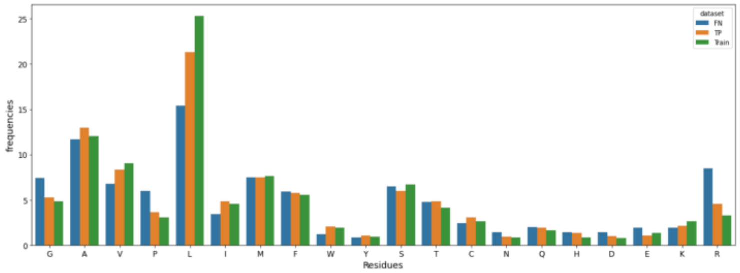


Sequence logo shows presence of leucine aa, indicating hydrophobicity and the peculiar A-X-A cleavage site in position 11-13.

The so-called AXA cleavage motif site is not conserved in the sequence logo of False Negatives [Fig.8. supplementary materials], where the alanine is substituted by valine and, the overall amino acidic composition is not the same with the training and true positives profiles [Fig.9. supplementary materials]. Since the amino acidic composition was the main feature to train the model, this explains the errors made during prediction.

In SVM, the assumption for discriminating between sequences having SP or not, relied on the sequence length and amino acidic composition. This because each sequence was encoded in a 20-d vector, considering up to k sequences positions. A possible way to understand if the principles were held also by the misclassified sequences, the amino acidic composition was computed for the training sequences and sequences classified as of true positives and false negatives in the benchmark set (Fig.10). The amino acidic composition of true positives is similar to the real positive composition of the sequences used to train the model. Regarding the AA composition of false negatives, the main differences can be noticed in leucine, which is less represented in the false negatives and arginine, which is in turn highly frequent in the false negatives and poorly represented in the positives.

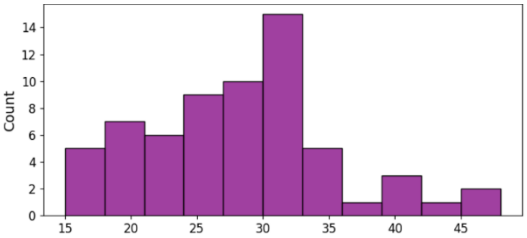
Fig.10. Residue frequency composition:



it's clear the difference in leucine and arginine in false negatives and real positive composition.

Furthermore, by analyzing the median and average length in sequence length distribution in true positive sequences in the benchmark and training sequences [Fig.11.supplementary materials], it is worth noticing that the average length for these sequences is between 20 and 25 for true positives of the benchmark and around 20 for positive training sequences [Fig.12.1 and Fig.12.2 supplementary materials]. Whereas, false negatives have a very different length distribution (Fig.13) respect to positive sequences and an average length above the selected hyperparameters k as it is shown in [Fig.14. supplementary materials]. The reason for the misclassification could be the result of the inclusion of some noise, for the smaller sequences; while, missing some information of the longer ones could not allow our classifier to exploit the whole region of interest, thus leading to errors.

Fig.13. False negatives SP length distribution**:**



Length distribution of sequences misclassified as negative in the benchmark set

Conclusions

Here, we propose two implemented machine learning algorithms, SVM and vonHeijne for signal peptide detection on amino acidic sequences. The 2 developed models were trained and tested using the same sets derived from SignalP-5 where the method's optimization was performed via cross validation. In SVM, cross validation procedure allowed the selection of the best parameters for the linear classifier, while in Von Heijne, cross-validation permitted to extract the optimal threshold for sequence classification in the test set.

Models are evaluated using five different parameters: accuracy, MCC score, precision, recall and F1 score. In the light of these values, SVM performed better than Von Heijne, in terms of sensitivity. However, our SVM model is outperformed by the latest released model for SP detection, Signal-IP 6.0.

Hence, False positive and False negative analysis were conducted to shed some light on the reasons behind these misclassifications. Sequences that have failed to be classified as SP might be related to the amino acid composition, which is not constant among all the sequences used to train both models. Also, some sequences display a more homogenous amino acidic composition. In addition, some sequences in the benchmark set have a length that falls shorter or longer with respect to the average length of training sequences. Moreover, a special treatment must be reserved for Transmembrane helices, which accounts for the highest false positive rates. These sequences should be treated properly in analysis of this kind.

References

Cai, Y.-D., Zhou, G.-P., & Chou, K.-C. (2003). Support Vector Machines for Predicting Membrane Protein Types by Using Functional Domain Composition. In *Biophysical Journal* (Vol. 84).

Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: A sequence logo generator. *Genome Research*, *14*(6), 1188–1190. https://doi.org/10.1101/gr.849004

Nielsen, H., Engelbrecht, J., & Brunak, S. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites Artificial neural networks have been used for many biological. In *Protein Engineering* (Vol. 10, Issue 1). http://www.tigr.org/.

Pedregosa FABIANPEDREGOSA, F., Michel, V., Grisel OLIVIERGRISEL, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot andÉdouardand, M., Duchesnay, andÉdouard, & Duchesnay EDOUARDDUCHESNAY, Fré. (2011). Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. In *Journal of Machine Learning Research* (Vol. 12). http://scikit-learn.sourceforge.net.

Petersen, T. N., Brunak, S., Von Heijne, G., & Nielsen, H. (2011). SignalP 4.0: Discriminating signal peptides from transmembrane regions. In *Nature Methods* (Vol. 8, Issue 10, pp. 785–786). https://doi.org/10.1038/nmeth.1701

Teufel, F., Almagro Armenteros, J. J., Johansen, A. R., Gíslason, M. H., Pihl, S. I., Tsirigos, K. D., Winther, O., Brunak, S., von Heijne, G., & Nielsen, H. (2022). SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nature Biotechnology*, *40*(7), 1023–1025. https://doi.org/10.1038/s41587-021-01156-3

VON HEIJNE, G. (1983). Patterns of Amino Acids near Signal-Sequence Cleavage Sites. *European Journal of Biochemistry*, *133*(1), 17–21. https://doi.org/10.1111/j.1432-1033.1983.tb07424.x

Von Heijne, G. (1986). A new method for predicting signal sequence cleavage sites. In *Nucleic Acids Research* (Vol. 14). https://academic.oup.com/nar/article/14/11/4683/2385409