Bioinformatics
doi: 10.1093/bioinformatics
Advance Access Publication Date: 25 Dec 2022
Manuscript Category

A comparison between different methods of Signal peptide prediction

Signal Peptide Prediction using VonHejne Method and Support Vector Machine

Omid Mokhtari[1]

[1]Department of Pharmacy and Biotechnology, University of Bologna, Italy

Abstract

Motivation: Detecting of the signal peptides in proteins can provide us new insights about the function and interactions of proteins that does not already have reliable experimental data. It could also lead to discovery of new potential drug targets. In this study, we build models using different methods based on support vector machine and position specific weight matrix and we compare their efficiency and potential failures.

Results: the support vector machine model that considered hydrophobicity proved to work better with the MCC and F1 score equal to 0.66 and 0.67 respectively.

Contact: omid.mokhtari@studio.unibo.it

Supplementary information: Supplementary data are available at https://github.com/amisteromid/Unibo-Projects

1 Introduction

The subcellular localization of proteins is strongly linked to their functional properties. Newly synthesized proteins destined for the secretory pathway, such as extracellular, periplasmic, or secreted proteins, use signal peptides to reach their final destination. These signal peptides are short sequences found in the N-terminal region of proteins and contain information for translocation. After translocation, a signal peptidase removes the signal sequence at the cleavage site.

At the sequence level, a complete signal peptide consists of three distinct regions: a hydrophobic core region surrounded by N-terminal and C-terminal regions. The N-terminal region is primarily composed of positively charged amino acids, while the C-terminal region contains the consensus recognition site for signal peptidase, A-X-A.

The detection of signal peptides and their cleavage sites is an important step in protein annotation, as it provides information on cellular components, which is one of the three aspects of Gene Ontology (GO). Additionally, by understanding protein localization, potential interacting proteins participating in the same biological process can be identified, and plasma membrane or cell surface proteins can be identified as potential drug targets.

To address this challenge, various algorithms have been developed. Gunnar von Heijne introduced the first method in 1983 based on a reduced-alphabet weight matrix, which aimed to recognize the cleavage site with the highest score. More recently, the sixth version of SignalP has demonstrated successful prediction of all types of signal peptides using language models in protein sequences. This study compares an updated version of the weight matrix approach and the support vector machine approach with various encodings. Both algorithms were trained using the dataset provided for SignalP 5.0.

2 Methods

2.1 Dataset Preparation

The present study utilized a dataset that was similar to the one used in SignalP-5.0 for training and benchmarking purposes. The data was derived from the UniProtKB 2018-04 release, which only included reviewed (SwissProt) sequences with a length greater than 30 residues. In addition, only signal peptides with experimental evidence for the cleavage site (ECO: 0000269) were taken into account. The N-terminal first 50 residues of the sequences were considered based on the potential location of the signal peptide.

The training set consisted of 1723 eukaryotic sequences, with 258 positive and 1465 negative samples. To ensure a fair evaluation, the dataset was randomly resampled into 5 equally-sized subsets, each with 345 sequences, while avoiding redundancy and similarity bias. The benchmark set was composed of 7456 sequences, including 209 positive and 7247 negative samples, and was derived through random selection from the original SignalP-5.0 dataset.

To compare the signal peptide length distributions in the training and benchmarking datasets, a density plot was generated (Fig. 1). A comparative analysis of the amino acid composition of the signal peptides in the benchmarking and training datasets with a random Swissprot background distribution was also performed (Supplementary Fig. 1). The results showed that the signal peptide length distribution was nearly identical, with a higher frequency of hydrophobic residues compared to the background composition in both the training and benchmarking datasets. This was further confirmed by the sequence logos (Supplementary Fig. 2), which displayed a common motif of AxA. Finally, pie plots were generated to present the distribution of proteins among species (Supplementary Fig. 3) and taxa (Supplementary Fig. 4).

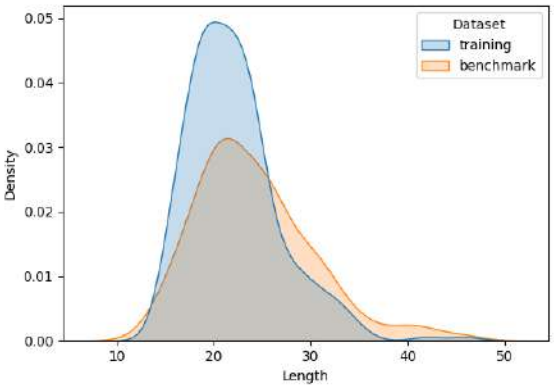

Figure 1 Distribution of signal peptide lengths in training and benchmarking dataset.

2.2 model evaluation scores

The performance of our models is regularly assessed using five statistical metrics, namely the Matthew's Correlation Coefficient (MCC), Accuracy, Precision, Recall, and F1 Score. The F1 Score is the harmonic mean of precision and recall, while MCC provides a more balanced and robust evaluation by taking into account all four categories of the confusion matrix.

		Predicted: Control	Predicted: Disease
Actual	Control	TN	FP
	Disease	FN	TP

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

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

2.3 VonHeijne method

Von Heijne proposed an algorithm designed for identifying signal peptides in proteins. The model focuses on the region surrounding the cleavage site, extending to the h-region. The training set consists of fragmented sequences (-13, +2), and a Position-Specific Probability Matrix (PSPM) is constructed to capture the frequency of each residue type at each position. To overcome the issue of zero probabilities in the PSPM, which would prevent computation of the log-odds, pseudo-counts are introduced during the computation. The PSPM is then used to generate the Position-Specific Weight Matrix (PSWM), which contains the log-odds between the frequencies in the PSPM and the Swissprot background distribution. Given a sequence X of length L, the log-likelihood score of X given the PSWM can be calculated as follows:

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

During the training procedure, we calculated the PSWM and obtained optimal threshold 5 times, each time with four folds, while the other fold was held out for cross validation. The final threshold was considered the mean between the 5 calculated thresholds. In each iteration, MCC, accuracy, precision, recall and F1 score were computed; the mean of results was used to describe the overall performance of training set. Finally, to run the algorithm with benchmark set, we built another PWSM using the whole dataset. Regarding the detection of signal peptide position, a sliding window of 15 residues was adopted to scan positions from 1 to 35 along protein sequence using PSWM scoring. The global score for the sequence was chosen to be the maximum one. Finally, the classification was done by evaluating the global score for each sequence with the threshold optimized during cross-validation.

2.4 Support Vector Machine

Support vector machines (SVM) are supervised machine learning methods used for classification. The algorithm is developed to separate two or more classes of points in a feature space, by maximizing the margin between the separating hyperplane and the datapoints with non-zero Lagrangian multiplier, known as support vectors. To allow soft margin classification, we introduced the hyper-parameter C to control the maximum margin and misclassification trade-off; high C value means high cost of error leading to perfectly separating points, and vice versa. Kernels are also adopted into in our model since classes could be not linearly separable; it manipulates the data by mapping it into another feature space with higher dimensions.

For detection of signal peptides, we implemented the SVM with radial-based function (RBF) kernel which computes the similarity. This kernel can be mathematically represented as:

$$K(X_1, X_2) = exp(-\frac{||X_1 - X_2||^2}{2\sigma^2})$$

Where, 'σ' (gamma) is the variance and our hyperparameter to be optimized; low values of gamma mean low curve of decision boundary, therefore broad decision region. Moreover, we need an estimation of signal peptide length to encode the first K residues of protein sequence. So, K is treated as another hyperparameter to be set for our final model.
Grid search is a technique to pick the best hyperparameters for fitting our model by evaluating all combinations of the sets of the values. For this purpose, we tried out the following combinations:

1) Values of K ranging between 20 and 24. The decision is made based on length distribution of signal peptides (fig.1)
2) Values of C ranging between 1 and 4
3) Gamma value set to 0.5, 1 or "scale" (scale is the default)

Regarding the encoding of protein sequences, we used three methods as described in the following:

- 20-dimensional vector corresponding to the normalized composition of the first K residues
- Position-based matrix of K*20 considering the frequency of each residue in specific positions
- 20+K-dimensional vector considering hydrophobicity of each position as an average hydrophobicity of a sliding window of size 5 based on Kyte and Doolittle scale. (the value for the first and last two residues are set as the residue hydrophobicity itself)

Optimal hyperparameter grid for each encoding method is set adopting 5-fold cross validation and MCC score. The average scores of folds are stored in table.1 for all three encoding methods in our SVM model. Finally, the testing was done on the benchmarking dataset using the selected parameters.

3 Results

In all methods, hyperparameters were set by cross validation and then performance of overall models was assessed by benchmark set.

3.1 VonHeijne results

5-fold cross validation in the VonHeijne method resulted in thresholds from 6.0 to 8.5. By making a profile from the whole training data and considering the average of 5 thresholds as our ultimate value, we obtained the performance rates reported in table 1. The number of false positives and false negatives were equal to 61 and 24 respectively (fig 2). Regarding the benchmark set, we had 277 and 37, false positives and false negatives respectively.

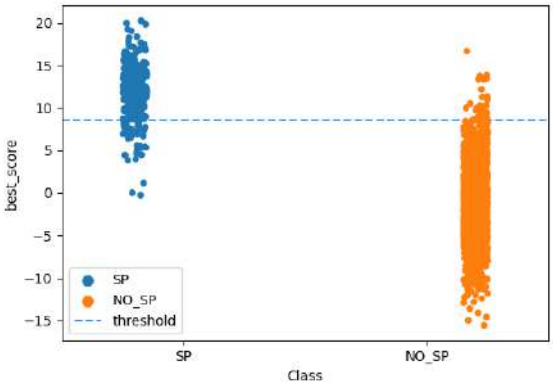

Figure 2 False negative, False positive trade-off

3.2 SVM results

After exploiting a grid search for the best possible combination for each encoding method, we came up by results reported in table 1. Optimum grid for k, gamma and c for different encoding was the following:

1st encoding) 20, scale, 2
2nd encoding) 22, scale, 4
3rd encoding) 22, scale, 1

However, performance rates for training set were almost the same but slight improvement in benchmark scores could be observed, as 0.67 F1 score was obtained in the third one.

Table 1 Training results

	Accuracy	Precision	Recall	MCC	F1 score
VonHeijne	0.950	0.793	0.906	0.819	0.846
SVM (1st encoding)	0.959	0.876	0.852	0.840	0.863
SVM (2nd encoding)	0.961	0.954	0.778	0.840	0.855
SVM (3rd encoding)	0.960	0.888	0.845	0.843	0.865

Table 2 Benchmarking results

	Accuracy	Precision	Recall	MCC	F1 score
VonHeijne	0.957	0.383	0.906	0.544	0.522
SVM (1st encoding)	0.973	0.525	0.746	0.613	0.616
SVM (2nd encoding)	0.981	0.669	0.650	0.650	0.660
SVM (3rd encoding)	0.980	0.620	0.736	0.666	0.673

4 Discussion

4.1 False positive analysis

Presence of common features among the false positives could lead us to the reason behind misclassification. Since signal peptides are composed of hydrophobic residues, other sequences with hydrophobic N-terminal like transit peptides and hydrophobic alpha helices could be mistaken as signal peptides. To investigate this possibility, we extracted proteins that have a transmembrane alpha helix or a transit peptide (either mitochondrion, chloroplast or peroxisome) in their N-terminal from the Uniprot database and then we used the number of negative results to calculate the feature-based FPR. In addition to the anticipated variations in the overall false positive rate (FPR) between the two approaches due to their respective efficiencies, the table demonstrates a significantly elevated transmembrane FPR when compared to other metrics. This outcome verifies that the model was primarily deceived by the occurrence of alpha helices, which should be addressed in order to enhance the model's performance.

Table 3 False positive rates calculated utilizing the comprehensive benchmark set, comprising of proteins with transmembrane elements (TM), transit peptides (TP), mitochondrial transit peptides (MTH), chloroplast transit peptides (CHP), and peroxisomal transit(PRX)

	FPR	TM	TP	MTH	CHP	PRX
VONHEIJNE	0.0380	0.3571	0.0590	0.0430	0.0355	0
SVM	0.0129	0.3285	0.0110	0.0422	0.0178	0

4.2 False negative analysis

Different analyses were performed for the false negatives of the two methods, as the errors might have arisen from the underlying assumptions in the modeling. For the von Heijne method, the focus was on the amino acid composition around the cleavage site as the main feature. To identify any differences in composition between the false negatives and true positives, their sequence logos were compared to that of the training set. As illustrated in the supplementary figure 5, the training set resembles the true positives more than the false negatives, with the latter showing a less prominent AXA cleavage site. This observation might highlights the need to incorporate additional features besides amino acid composition to improve the accuracy of the model.

To assess the errors in the support vector machine method, the length distribution of the signal peptides for both false negatives and true positives was plotted and analyzed (as seen in supplementary figure 6). The results indicated that many of the false negatives had a length that deviated from the chosen k parameter. This suggests that there may have been some noise introduced for signal peptides shorter than k and that the full signal peptide may not have been captured for those longer than k. Further examination was conducted by comparing the amino acid composition of the false negatives to that of the true positives and the training set. The comparison (shown in supplementary figure 7) revealed disparities in the false negatives' composition, with a higher prevalence of arginine and a lower presence of leucine. Finally, to determine if these differences were due to a distinct species distribution among the false negatives, a comparison was made with the other sets, but no significant differences were detected.

5 Conclusion

This project aimed to construct a model that can predict the existence of signal peptides in protein sequences. Two machine learning algorithms, the von Heijne method and the support vector machine, were utilized for comparison. Both methods were trained through 5-fold cross validation, utilizing the SignalP 5.0 training dataset, which offered an adequate number of sequences to optimize the relevant hyperparameters. The models' performance was evaluated on the benchmark set using metrics such as MCC, accuracy, precision, recall, and F1 score. The support vector machine was found to be more efficient, with a superior overall performance compared to the von Heijne method. However, both models produced numerous misclassifications during the testing phase. The dataset was then analyzed to determine the causes of the errors for both methods. The analysis revealed that sequences without signal peptides that were classified as positive (FP) had transit peptides and, more notably, transmembrane alpha helices, which may have led to misclassification as these structures share a hydrophobic composition similar to that of SPs. For false negatives, a comparison based on amino acid composition and length distribution was made with false positives and the training set to identify any differences. False negatives were found to have different SP lengths and amino acid compositions (more valine and arginine instead of leucine and alanine). From these results, it can be concluded that most of the errors were due to limitations in the algorithms' training and that improving the model's performance may be possible by incorporating different features.

References

von Heijne, G. (1990). The signal peptide. The Journal of membrane biology, 115, 195-201.

Inouye, S., Soberon, X., Franceschini, T., Nakamura, K., Itakura, K., & Inouye, M. (1982). Role of positive charge on the amino-terminal region of the signal peptide in protein secretion across the membrane. Proceedings of the National Academy of Sciences, 79(11), 3438-3441.

Owji, H., Nezafat, N., Negahdaripour, M., Hajiebrahimi, A., & Ghasemi, Y. (2018). A comprehensive review of signal peptides: Structure, roles, and applications. European journal of cell biology, 97(6), 422-441.

Nielsen, H., Brunak, S., & von Heijne, G. (1999). Machine learning approaches for the prediction of signal peptides and other protein sorting signals. Protein engineering, 12(1), 3-9.

Almagro Armenteros, J. J., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., ... & Nielsen, H. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. Nature biotechnology, 37(4), 420-423.

Noble, W. S. (2006). What is a support vector machine?. Nature biotechnology, 24(12), 1565-1567.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics, 21, 1-13.

Chou, K. C. (2001). Prediction of signal peptides using scaled window. peptides, 22(12), 1973-1979.