Sequence analysis

Evaluating signal peptide prediction in eukaryotes: support vector machine VS position-specific weight matrix approach

Irene D'Onofrio[1, *]

[1]Department of Pharmacy and Biotechnology, University of Bologna, 40126 Bologna, Italy

*To whom correspondence should be addressed.

Abstract

Motivation: Signal peptides (SP) act as molecular 'zip codes' that guide the proteins to the secretory pathway. Given the critical significance of signal peptides in elucidating protein roles and their relevance in therapeutic applications, there is a compelling need to develop methodologies for the identification of signal peptides.

Results: Here, we introduce a Support Vector Machine (SVM) model that incorporates several features to account for both SP characteristics and the broader protein context in eukaryotic proteins. The SVM model is then compared to the weight matrix-based approach that traces the method proposed by Gunnar von Heijne in 1986. When tested on a blind set, the weight matrix-based approach yields an MCC of 0.70, while the SVM resulted in a much more precise, sensitive, and performing prediction method with an MCC of 0.89.

Contact: irene.donofrio@studio.unibo.it

1 Introduction

Comprehending the subcellular distribution of proteins plays a pivotal role in unraveling their functional roles. Secreted proteins, collectively constituting the secretome, undergo active transport via the secretory pathway, ultimately exiting the cell. In eukaryotes, proteins pass through the ER, Golgi apparatus, and vesicles before reaching the plasma membrane for release. In prokaryotes, translocation happens across the cytoplasmic membrane. Both eukaryotes and prokaryotes rely on a short peptide, the Signal Peptide (SP) (Von Heijne, 1990) to target the proteins to the secretory pathway. During the translocation process the SP is cleaved off and most of the signal peptides are removed by signal peptidase (SPase) I (LepB in Bacteria), which has orthologs in Archaea and Eukarya (Dalbey et al., 1997).

A typical SP has 15–30 residues and is characterized by a tripartite structure: i) N-region: the positive-charged domain (1-5 residues), ii) H-region: the hydrophobic core (7-15 residues), and iii) C-region: region flanking the cleavage site (3-7 residues), as shown in Figure 1. The hydrophobic core (H-region) is the cardinal element of the SP. It determines the conformation of SP, the protein processing, the cleavage, and the secretory pathway (Sec, SRP o Tat). This region, which tends to form an alpha helix, is formed of at least 7 residues. Leu predominates in both prokaryotes and eukaryotes.

Generally, SPs show a great variability, however, they seem more conserved around the cleavage site, indeed, positions - 1 and - 3 seem to be strongly selected for small and neutral residues (Von Heijne, 1983). This conservation is strict in prokaryotes, unlike eukaryotes (Choo and Ranganathan, 2008).

The majority of SPs that are transported via Sec/Tat/SRP pathways in all life domains have the above-mentioned tripartite structure, however, there are other types of structures. Twin arginine (RR)-translocated SPs feature a twin-arginine motif, while SPs cleaved by SPase II feature a C-terminal lipobox. Sec/SPIII SPs have no substructure (Dalbey et al., 2012).

Signal peptides play an important role in a variety of applications, from recombinant protein production to disease diagnosis and vaccination (Ohmuro-Matsuyama and Yamaji, 2018; Kovjazin et al., 2011). Hence, it is plausible the interest within the scientific community for the development of signal peptide prediction tools.

Signal peptide prediction comprises a classification and a labeling task, where the former involves the discrimination of SP-proteins from non-SP protein, whereas the latter the prediction of the position of the cleavage site. However, these tasks come with some challenges. One of the main difficulties is the distinction between SPs and N-terminal transmembrane helices or transit peptides, united by the presence of hydrophobic regions, even though of different length. As for the prediction of the cleavage site, the complexity derives from the sequence variability and the lack of a strong conservation, especially in eukaryotes.

The SP prediction is one of the earliest challenges in the bioinformatics field, the first attempt was introduced in 1983 (Von Heijne, 1983). Since then there has been a continuous improvement, first with the use of weight matrices (von Heijne, 1986), and then with ANNs (Nielsen et al., 1997).

Figure 1. General structure of a SP. It is composed of three main parts: 1) N-region- the positive-charged domain 2) H-region- the hydrophobic core, that tends to form an α-helix 3) C-region- the cleavage site.

Subsequently, the usage of HMMs was explored to model the length difference between SP and other N-terminal hydrophobic structures, such as signal anchors (Nielsen and Krogh). As for SVMs, one of the most successful results was obtained by Vert, who trained an SVM for SP cleavage sites using a new class of kernels for strings (Vert, 2002). More recently better results were achieved applying deep learning techniques based on convolutional ANNs (Savojardo *et al.*, 2018) and recurrent ANNs (Almagro Armenteros *et al.*, 2019). In 2022, the latest version of SignalP was published (Teufel *et al.*, 2022), a method based on protein language models able to identify all five types of SPs, with a performance superior to most of the existing predictors.

In this study, a Support Vector Machine (SVM) is implemented for the specific prediction of SPs in eukaryotic proteins. This model incorporates several features that encode characteristics of both the SP sequence itself and of the entire protein, resulting in a more comprehensive predictive model. When compared to a position-specific weight matrix (PSWM) method, that traces one of the earliest approaches (von Heijne, 1986), the SVM outperformed it. Indeed, when benchmarked on the same blind set as the weight matrix-based method, the SVM revealed to be a better performing method, with an MCC of 0.89, higher sensitivity and precision.

2 Methods

2.1 Datasets

2.1.1 Data collection

The data, comprising only eukaryotic proteins, were retrieved from UniProtKB/SwissProt release 2023_04 (The UniProt Consortium, 2023).

The positive set (SP proteins) comprises 2942 entries, which specifically contains proteins with experimentally annotated signal peptide cleavage sites. Proteins with signal cleavage sites that were either unclear or positioned at a site earlier than the 13th position were excluded from this set.

The negative set (non-SP proteins) comprises 30011 entries that lack a signal peptide at any evidence level and are annotated with experimental evidence for the following cellular compartments unrelated to the signal peptides: cytosol, nucleus, mitochondrion, plastid, peroxisome, cell membrane. A further filtering was carried out to not include in the set the proteins containing terms related to the secretory pathway (i.e., 'endoplasmatic', 'golgi', 'secreted' and 'lysosome'). Both positive and negative proteins are longer than 30 residues.

2.1.2 Training and benchmarking datasets generation

In order to reduce the redundancy between the training and the benchmarking datasets and permit an unbiased evaluation of the outcomes, MMseq2 (Steinegger and Söding, 2017) was adopted using a sequence identity threshold of 30%, a pairwise alignment coverage of at least 40% and a connected component strategy. This resulted in 1093 representatives for the positive dataset and 9523 representatives for the negative dataset. The data was divided into two sets: a training dataset, which comprises 80% of the data, and a benchmarking dataset (i.e., blind test set), which consists of the remaining 20%. To maintain the proportion of positive and negative entries, the 80-20% split was independently applied to both the positive and negative datasets (Table 1). Additionally, while preserving the positive-negative ratio, the training dataset was divided into 5 subsets to be adopted in a 5-fold cross-validation process. The similar distribution of the SP lengths among the 5 subsets demonstrates that the partitioning was balanced (Supplementary Figure S1). The complete dataset is available in the Supplementary.

Table 1. The dataset adopted in this work.

Dataset	No. of SP proteins	No. of non-SP proteins	Total
Training	874	7618	8492
Benchmarking	219	1905	2124
Total	1093	9523	10616

2.1.3 Transmembrane and transit peptide annotations

For the negative entries belonging to the benchmarking dataset, feature annotations about the presence of a transmembrane region and/or a transit peptide were retrieved from UniProtKB/SwissProt. The entries were labeled as 'Transmembrane Proteins' (TM) or as 'Protein with a Transit Peptide' (TS) exclusively if the annotations were manually curated or there was experimental evidence (corresponding to the UniProtKB evidence codes ECO: 0000269, ECO: 0000305, ECO:0000250, ECO:0000255, ECO:0000312, ECO:0007744). Another required condition was that at least one TM region/TS region was present in the first 90 residues.

2.1.4 Statistical analyses

Statistical analyses were conducted to evaluate whether there are any significant distinctions between the training and blind test sets and the positive and the negative sets, and to identify unique aspects of the signal peptides. These analyses focused on examining the length of signal peptides (Fig. 2a) and of the proteins (Fig. 2b), the N-terminal residue composition compared to a background distribution (Fig. 2e), the sequence logo of the cleavage sites (Fig. 2c-d) and the taxonomic distribution both at the kingdom and at the species level (Supplementary Figure S2-3-4).

The analysis of the signal peptide lengths allowed to identify the minimum, maximum and average lengths for the signal peptides, which are 13, 64 and ~23, respectively (Fig. 2a). Whereas the comparison of the residue composition between the signal peptides against a background distribution (represented by eukaryotic proteins in UniProtKB-SwissProt) revealed an enrichment in alanine, leucine, methionine, valine, and phenylalanine within the signal peptide sequences (Fig. 2e). The hydrophobic nature of the latter four residues aligns with the presence of the characteristic hydrophobic core typically found in signal peptides. It is important to consider that the enrichment of methionine may be influenced by the fact that the signal peptide corresponds to the N-terminal region. The sequence logos of the SP cleavage sites for both training and benchmarking datasets were displayed using WebLogo (Crooks et al., 2004; Schneider and Stephens, 1990). From the logos it is possible to observe the hydrophobic core and the light conservation in position -1 and -3 of the AXA (VXA) motif (Perlman and Halvorson, 1983; Von Heijne, 1983), whereas positions +1, -2, and -4 showed a low information content, indicating that they can accommodate almost any kind of residue (Fig. 2c-d). It has to be considered that in eukaryotic organisms, the conservation of this motif is relatively weaker compared to prokaryotes (Choo and Ranganathan, 2008), as mentioned above.

As for what concerns the comparison between the SP and non-SP proteins, minor differences were observed in the distribution of the protein length, with negative entries being slightly longer (Fig. 2b). However, disparities emerged in terms of taxonomy. The positive set showed a much more asymmetric distribution of both kingdom and top 10 species (Supplementary Figure S2-3-4). Metazoa stood out as the most prominent kingdom in

the positive class, with *Homo sapiens* being the most represented species. This discrepancy could be due to the fact that Metazoa SP proteins, and especially human ones, are more studied and thus more annotated in UniProtKB/Swiss-Prot. Furthermore, an analysis of the global residue composition between SP and non-SP proteins revealed some differences (Supplementary Figure S5). SP proteins exhibit a higher frequency of apolar residues, whereas non-SP proteins exhibit a higher frequency of charged residues. The cysteine, contributing to protein stability, is quite represented in proteins that are involved in the secretory pathway, making it more abundant in SP proteins (Robinson and Bulleid, 2020). Regarding the comparison between the training and benchmarking sets, an examination of the signal peptide length distribution revealed that there are no notable distinctions (Fig. 2a). While the benchmarking set exhibits slightly greater variability, this can be attributed to its smaller size. Additionally, when considering the distribution of protein length (Fig. 2b), residue composition (Fig. 2e), taxonomic distribution (Supplementary Figure S2-3) and the sequence logos of the cleavage sites (Fig. 2c-d), both sets exhibit no discernible distinctions, signifying that the process of 80-20% splitting was executed fairly.

2.2 von Heijne method

2.2.1 General idea and description of the algorithm

The von Heijne method (von Heijne, 1986) falls within the weight-matrix methods. Firstly, the frequencies of each kind of residue in each position in a sample of aligned sequences (in this case the region around the cleavage site, -13 +2) are computed, resulting in the PSPM (Position Specific Probability Matrix). Subsequently, the PSPM is normalized by dividing the frequencies for each residue in each position by its respective relative abundance in a background model and computing the natural logarithm of this ratio, this leads to the PSWM (Position Specific Weight Matrix). The PSWM, being normalized by a background model, allows to highlight possible over/under-representations. Any novel sequences can be then scanned by a sliding window; this process yields a quantitative assessment of how well the sequence aligns with the reference sample adopted in constructing the weight matrix. The window with the highest score will give

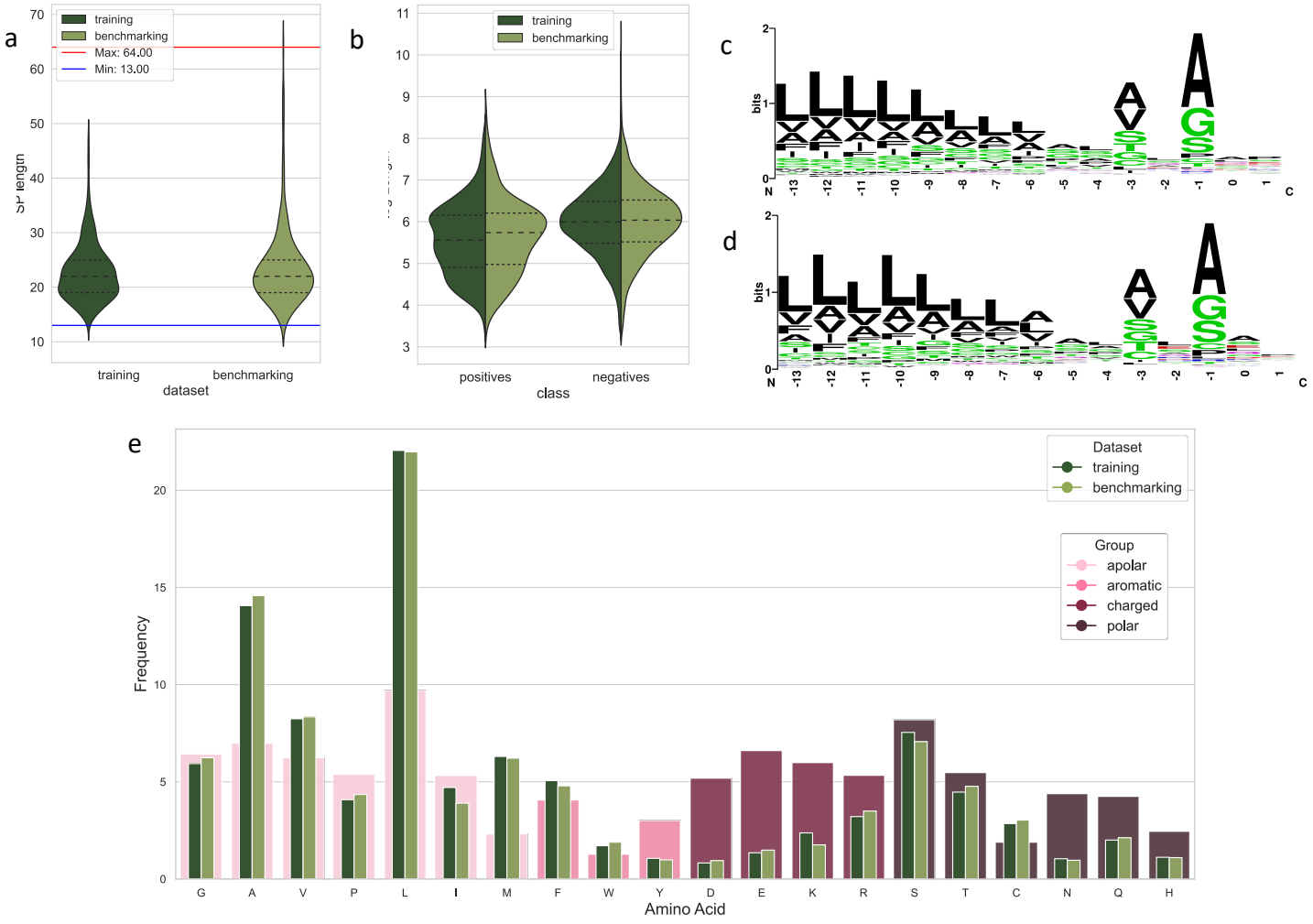

Figure 2. Statistical analyses of the data. **a.** signal peptide length distribution within the training and benchmarking datasets; maximum and minimum length are highlighted. **b.** proteins log length distribution in training and benchmarking and SP and non-SP proteins. **c.** sequence logo of the -13 +2 region of the sequences belonging to the training dataset; the total height of the stack of letters at each position shows the amount of information, while the relative height of each letter shows the relative abundance of the corresponding amino acid; uncharged polar residues are green and hydrophobic residues are black. **d.** sequence logo of the -13 +2 region of the sequences belonging to the benchmarking dataset. **e.** residue composition of the SP sequences in both training and benchmarking datasets, compared with a background distribution constituted by eukaryotic proteins in UniProtKB/SwissProt.

the position of the putative cleavage site. This approach not only allows for the detection of the most probable signal peptide, but also permits a discrimination between signal peptide sequences and the N-terminal region of proteins not involved in the secretory pathway. About the work presented here, the original von Heijne method was replicated with a few changes and aimed solely at the classification of proteins with signal peptides.

Given a set of N stacked sequences of length L (L=15), the PSPM was filled as follows:

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

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; $I(s_{i,j} = k)$ is an indicator function (1 if the condition is met, 0 otherwise). Pseudocounts (+1) are added to avoid zero values, thus the observations were divided by N+20, to account for pseudocounts. To compute the PSWM, the values are corrected for the background frequency in UniProtKB-SwissProt and the logarithm of the ratio computed:

$$W_{k,j} = log\frac{M_{i,j}}{b_k} \quad (2)$$

Once the PSWM is built, any sequences can be scanned by a sliding window of length 15 (length of the PSWM), until the 75th residue (90-15). For each window $X(X_1, \dots, X_{15})$ the score is computed as**:**

$$score_{(X|W)} = \sum_{i=1}^{L} W_{x_i i} \quad (3)$$

Then for each sequence the maximum score among all the window scores is picked, the higher the score the higher the probability that the correlated sequence is a signal peptide. To classify the entries, a threshold must be chosen: the entries with a score lower than the threshold will be classified as non-SP proteins, vice versa the entries with a score greater than or equal to the threshold will be classified as SP proteins.

2.2.2 Threshold optimization (CV), training and benchmarking

To define an optimal threshold, a 5-fold cross-validation procedure was implemented, ensuring that all the subsets (see Section 2.1.2) had the opportunity to explore each role. Each run is characterized by the following steps:

- Training: three subsets are adopted to compute the PSWM
- Validation: one subset is adopted to pick the threshold that maximizes the F1-score.
- Testing: the entries of the remaining subset are scored with the PSWM and then classified using the optimal threshold for that run. The performance is evaluated by adopting different metrics.

Precision-Recall curves and ROC curves of the cross-validation are in the Supplementary (Supplementary Figure S6a-b).

At the end of the cross-validation, the optimal threshold to use in the benchmarking phase was obtained by averaging the 5 thresholds obtained from the individual runs. Upon completing the threshold optimization procedure, the PSWM was computed on the entire training dataset (Supplementary figure S7) and the performance benchmarked on the blind set.

2.3 SVM

2.3.1 Support Vector Machine for classification task

Support Vector Machines (SVM) are supervised learning methods used both for regression and classification tasks, proposed by Vapnik and coworkers in 1992 (Boser et al., 1992). As for classification tasks, the goal of SVM is to predict the target values of the test data given only the test data attributes. Given a set Ω of vectors ($i = 1, \dots, m$) partitioned into two classes, vectors are represented by a pair (x_i, y_i) $\in \mathbb{R}^n$ x $\{-1, 1\}$, where n is the number of features observed for each vector, x_i contains the feature values for vector i and y_i indicates to which of the two classes of Ω vector i belongs. Support Vector Machines look for the hyperplane $f(x) = w^T x + b$ that maximizes the distance (or margin) $2/\|w\|$ between two parallel hyperplanes ($w^T x$ + b = 1 and $w^T x$ + b = −1) supporting some vectors of the two classes and that maximizes the sum of the misclassification error. Thus, the constraint-optimization problem of the SVMs is as follows:

$$\min_{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \quad (4)$$

$$s.t \quad y_i\,(w^T x_i + b) \geq 1 - \xi_i \qquad i = 1, \dots, m \quad (5)$$

$$\xi_i \geq 0 \qquad i = 1, \dots, m \quad (6)$$

The objective function (4) consists in the maximization of the margin the minimization the sum of the misclassification (indicated by the slack variable ξ_i). The second term is multiplied by C which is an hyperparameter that regulates the trade-off between the two objectives. In brief, C determines the softness of the margin: the lower the value of C, the lower the penalty and the larger the margin hyperplane (soft margin), whereas the higher the value of C, the higher the penalty and the smaller the margin hyperplane (hard margin). The constraints (5) ensures that each example must be correctly classified by the decision boundary (y_i ($w^T x_i$ + b) ≥ 1) or, if it is misclassified, the misclassification error should be small, as indicated by the slack variable ξ_i. The constraint (6) ensures that all the slack variables are greater than or equal to 0. Eventually, the function optimized (maximized) in the SVMs is the Dual Lagrangian:

$$\tilde{\mathcal{L}}(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j\, x_i^T x_j \quad (7)$$

Subject to the constraints:

$$0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{n} y_i\alpha_i \quad (8)$$

The geometrical interpretation is depicted in Figure 3a.

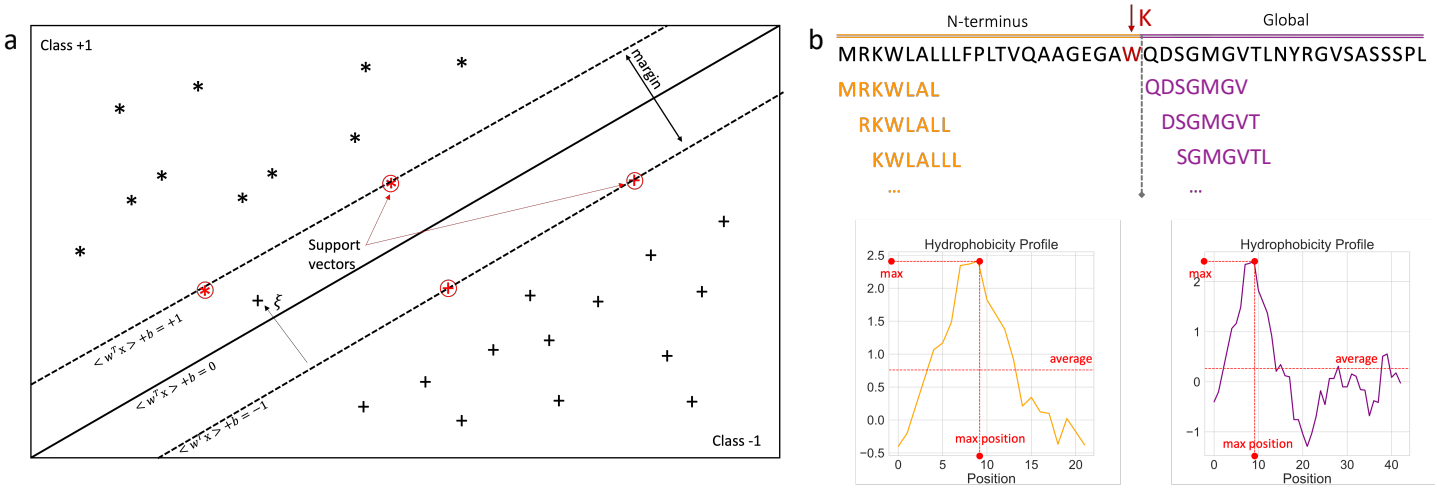

Figure 3. **a**. Geometrical representation of the SVM method. **b**. Feature extraction using a sliding window of length 7 until K (optimized hyperparameter) and from K until the end. For each feature, the max value, max value position and the average value are encoded. In the image above, the hydrophobicity feature is taken as example.

To deal with not linearly separable classes, it is possible to adopt the Kernel trick, which consists in mapping the vectors x_i into a higher (maybe infinite) dimensional space by the function φ. SVM finds a linear separating hyperplane with the maximal margin in this higher dimensional space. The kernel function is defined as follows:

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j) \quad (9)$$

One of the most common kernel functions is the Gaussian radial basis function (RBF):

$$K(x_i, x_j) = exp\left(-\gamma \|x_i - x_j\|^2\right), \quad \gamma > 0 \quad (10)$$

RBF is a function whose value depends on the distance (usually Euclidean distance) between two examples (x_i and x_j) in the input space. The hyperparameter γ controls how far or how little the influence of each datapoint is felt. In this work the RBF is adopted.

2.3.2 Application of SVM to SP prediction

For SVM, the issue of SP/Non-SP discrimination is a two-class problem to distinguish proteins with signal peptides from proteins with no signal peptides.
Each sequence is encoded as a vector, whose dimension depends on the number of features selected. As a baseline vector, a 20-dimensional vector containing the frequency of each residue at the N-terminus ('Ncomp') is adopted. This baseline vector can be extended with the following additional features:

- Global composition ('gcomp'): frequency of each residue starting from the Kth position until the end.
- Hydrophobicity ('hp'): this feature has been extracted using a sliding window of 7 until K and the hydropathicity scale of Kyte and Doolittle (Kyte and Doolittle, 1982).
- Global hydrophobicity ('ghp'): the profile is computed adopting a sliding window of 7 and the same propensity scale adopted for 'hp' (Kyte and Doolittle, 1982).
- Charge ('ch') = the scale is structured such that positively charged residues have value 1 and the other residues have a value equal to 0. The sliding window adopted for the extraction was of length equal to 5.
- Helix tendency ('ht'): the profile is computed adopting a sliding window of 7 that slides until K, the scale adopted is the one developed by Chou and Fasman (Chou and Fasman, 1978).
- Transmembrane tendency ('tmt'): it is encoded using the Zhao and London scale (Zhao and London, 2006) and a sliding window of length 7.

For all these features but global composition, the average value, the maximum value, and the position of the maximum value are extracted (see Figure 3b) and normalized within the range [0,1].

2.3.3 Hyperparameters optimization and model selection

There are three parameters that need to be optimized in a grid search within a cross-validation procedure:

- C: controls the tradeoff between the maximization of the margin and the minimization of the possibility of misclassification errors.
- γ: defines how far the influence of a single training example reaches.
- K: position until/since the features are tested. Its value range around the average SP length.

Table 2. Results of the cross-validation procedure for all the models. Each metrics is accompanied by the standard error. *Note:* the complete table with the hyperparameters can be found in the Supplementary

Model	ACC	MCC	Precision	Recall	F1-score
von Heijne	0.94 ± 0.00	0.67 ± 0.01	0.70 ± 0.04	0.72 ± 0.04	0.70 ± 0.01
SVM	0.96 ± 0.00	0.77 ± 0.02	0.83 ± 0.02	0.75 ± 0.03	0.78 ± 0.03
SVM-gcomp	0.97 ± 0.00	0.84 ± 0.02	0.89 ± 0.01	0.82 ± 0.02	0.85 ± 0.02
SVM-hp	0.97 ± 0.00	0.83 ± 0.02	0.86 ± 0.01	0.83 ± 0.03	0.84 ± 0.02
SVM-ghp	0.96 ± 0.00	0.78 ± 0.02	0.85 ± 0.01	0.77 ± 0.03	0.80 ± 0.02
SVM-ch	0.96 ± 0.00	0.78 ± 0.02	0.83 ± 0.02	0.78 ± 0.02	0.80 ± 0.02
SVM-ht	0.96 ± 0.00	0.78 ± 0.02	0.83 ± 0.01	0.78 ± 0.02	0.81 ± 0.02
SVM-tmt	0.97 ± 0.00	0.85 ± 0.01	0.87 ± 0.01	0.86 ± 0.01	0.86 ± 0.01
SVM-tmt-gcomp	0.98 ± 0.00	0.87 ± 0.01	0.90 ± 0.01	0.87 ± 0.02	0.89 ± 0.01
SVM-tmt-hp	0.97 ± 0.00	0.85 ± 0.01	0.86 ± 0.01	0.87 ± 0.02	0.86 ± 0.01
SVM-tmt-ghp	0.97 ± 0.00	0.85 ± 0.01	0.88 ± 0.01	0.86 ± 0.01	0.87 ± 0.01
SVM-tmt-ch	0.97 ± 0.00	0.86 ± 0.01	0.87 ± 0.01	0.87 ± 0.01	0.87 ± 0.01
SVM-tmt-ht	0.97 ± 0.00	0.85 ± 0.01	0.86 ± 0.00	0.86 ± 0.01	0.86 ± 0.01
SVM-tmt-gcomp-hp	0.98 ± 0.00	0.89 ± 0.00	0.91 ± 0.01	0.89 ± 0.00	0.90 ± 0.00
SVM-tmt-gcomp-ghp	0.98 ± 0.00	0.89 ± 0.00	0.91 ± 0.01	0.89 ± 0.01	0.90 ± 0.00
SVM-tmt-gcomp-ch	0.98 ± 0.00	0.88 ± 0.01	0.90 ± 0.01	0.89 ± 0.01	0.89 ± 0.01
SVM-tmt-gcomp-ht	0.98 ± 0.00	0.88 ± 0.01	0.91 ± 0.00	0.88 ± 0.01	0.89 ± 0.00

A complete list of hyperparameters tested and their optimal values are available in Supplementary Table S2.

For each model a 5-fold cross-validation was implemented as described for the von Heijne method (see Section 2.2.2). Within each run of the cross-validation, all the possible combinations of the hyperparameters were exploited in the training phase and the combination that maximized in the validation subset the MCC was picked to be then tested. At the end of the cross-validation procedure, for each model the chosen combination of the hyperparameters includes the most frequent value of each hyperparameter. In case of tie, for C and gamma the lowest value is picked, whereas for K the value that is closest to the average length of the signal peptides.

Regarding the model construction process, it commences with the base model, which consists solely of a 20-dimensional vector representing residue frequencies up to K. Each feature is subsequently introduced individually, and the performance of the resulting feature-enhanced model is assessed. The new base model is then chosen based on its ability to maximize the MCC. Then, the other features are individually added to the new base model, only if they improved or maintained the same performance once added to the previous base model. The iterative process persists until there is no longer a significant improvement in the MCC value, signifying that the model has reached its optimal configuration.

2.4 Performance metrics

To assess the models' performances, the following metrics were used:

- Accuracy (ACC):

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \quad (11)$$

- Matthews correlation coefficient (MCC):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (12)$$

- Precision:

$$Precision = \frac{TP}{TP + FP} \quad (13)$$

- Recall (TPR):

$$Recall = TPR = \frac{TP}{TP + FN} \quad (14)$$

- F1 Score (harmonic mean of precision and recall):

$$F1\ score = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (15)$$

- FPR (False Positive Rate):

$$FPR = \frac{TP}{TP + FP} \quad (16)$$

The FPR was also computed with respect to the transmembrane proteins (TM) and the transit peptide (TS). In this specific case the rate derives from the ratio of the misclassified TM and TS as signal peptide (TP_{TM} and TP_{TS}) over the total number of TM or TS among the negative entries of the benchmarking dataset, respectively.

- FNR (False Negative Rate):

$$FNR = \frac{FN}{TP + FN} = 1 - TPR \quad (17)$$

3 Results

3.1 CV results

The results of the 5-fold cross validation procedure for both the von Heijne method and the SVM models are reported in Table 2. Each metric derives from the mean of the results of the 5 runs and is accompanied by the standard error. The baseline SVM model, which incorporates the 20-dim vector with the N-terminal frequency of each residue, is referred simply as 'SVM'.

The 5-fold cross validation procedure for the von Heijne resulted in an optimal threshold equal to 9.06 ± 0.31, with a corresponding MCC in the validation step of 0.67 ± 0.01. Even the base SVM model outperformed the weight-matrix based approach, achieving an MCC of 0.77± 0.02.

The models have been constructed as described in the Section 2.3.3 With respect to the base model, the first feature that augments the performance is the transmembrane tendency (MCC: 0.85 ± 0.01), which not only considers the hydrophobicity, but also could aid in the identification of N-terminal transmembrane regions. Subsequently, the global composition emerges as the next influential feature, resulting in an MCC of 0.87 ± 0.01. It is followed closely by either the global hydrophobicity feature or the N-terminal hydrophobicity feature, both of which produce an MCC of 0.89 ± 0.01.

Comparing the results in Table 2, two models exhibit equally strong performances in the classification task: 'SVM-tmt-gcomp-hp' and 'SVM-tmt-gcomp-ghp'. The only difference between these two models lies in only one feature, the hydrophobicity, which is local (until K) in one case and global in the other.

The model chosen to be benchmarked is 'SVM-tmt-gcomp-ghp'. This choice is motivated by the fact that this model encompasses both local (transmembrane tendency) and global (hp global) scales to account for hydrophobicity. This combination should enhance the comprehensiveness of the model.

3.2 Benchmarking

The von Heijne method and the top performing SVM model were subjected to benchmarking using the blind set. Notably, both methods exhibit comparable performance in both cross-validation and the blind test, suggesting their generalization ability and ruling out overfitting concerns. The von Heijne reaches an MCC of 0.70, slightly higher than the MCC of the validation.

However, as in cross-validation, the best SVM model outperformes the von Heijne method, yielding an MCC of 0.89 (Table 3).

Also, looking at the other parameters, the SVM model is more precise than sensitive, unlike the von Heijne method.

The confusion matrices of both methods are in the Supplementary material (Supplementary Figure S8).

Table 3. Benchmarking results

Model	ACC	MCC	Precision	Recall
von Heijne	0.94	0.70	0.66	0.82
SVM	0.98	0.89	0.93	0.87

3.3 False positive and negatives analysis

3.3.1 FP analysis

As illustrated in Table 4, the False Positive Rate (FPR) exhibited a notable reduction in the case of the SVM model.

Nevertheless, the presence of an FPR percentage for the SVM model, albeit relatively small, indicates the necessity for a more in-depth analysis of false positives.

The reason behind these classification errors could reside in the hydrophobicity core that characterizes the signal peptides. Other hydrophobic elements similar to SP usually present at the N-terminus, such as transit peptides and transmembrane proteins, could be misclassified. The presence of these two types of proteins within the false positives was investigated, the data for such analyses were retrieved as described in Section 2.1.3.

The notable reduction for the SVM classifier with respect to the von Heijne method in error rates extends to False Positive Rates (FPR) associated with transit peptides (FPR_{TS}) and especially transmembrane proteins (FPR_{TM}), for which the reduction is of 22% circa. The TM and TS in the benchmarking set are 149 and 216, respectively.

However, as displayed in Supplementary Figure S9, both the von Heijne method and the SVM model exhibit a statistically significant enrichment of transmembrane proteins within their FP predictions, compared to TN (VH-Fisher's exact P=5.90e-22, SVM-Fisher's exact P=5.88e-5). In the case of transit peptides, the von Heijne method also shows an enrichment within the false positives (Fisher's exact P=0.0439), while the SVM classifier showed no significant difference in proportions (see Supplementary Figure S9).

Table 4. False positive rates

Model	FPR (%)	FPR_{TM} (%)	FPR_{TS} (%)
von Heijne	4.9	26.9	7.9
SVM	0.8	4.7	0.5

3.3.2 False negatives analysis

The comparison of False Negative Rate (FNR) between the two models reveals a reduction for the SVM classifier (FNR-VH=17.8%, FNR-SVM = 12,8%), which indicates an improvement in the sensitivity.

As for the von Heijne method, the presence of FN could be due to a different composition of those entries around the cleavage site. To further analyze this hypothesis, the sequence logo of the false negatives is displayed (see Supplementary figure S10). Over all the length analyzed, the FN sequence logo shows less information, compared to the sequence logo of the benchmarking dataset (Fig. 2d). Indeed, from position -13 to -6, the typical abundance of leucine, alanine and valine is not evident. The same can be stated for the positions -3 -1 about the AXA motif.

Regarding the SVM classifier, the misclassification could be attributed to a different composition of the N-terminus, which is the base feature with which the classifier has been implemented. Additionally, the presence of false negatives could be influenced by the length of the signal peptide. The SVM classifier incorporates the K parameter, which is set close to the average length of signal peptides, however, among the false negatives, there might be proteins with signal peptide lengths that deviate from this mean length. To further explore these hypotheses, both the N-terminal residue composition (until residue 22) and the distribution of the SP length are examined (Supplementary figures S11-12). The N-terminal residue composition of the FN and TP doesn't show a neat difference, however some frequencies are dissimilar; leucine, alanine and valine are less represented in the FN, while other residues, such as arginine, unexpectedly appear more abundant among the FN cases. As for the SP length, quite clearly the distribution of the FN was more stretched, with the minimum and maximum lengths extending to more extreme values.

4 Conclusion

In this work, a support vector machine (SVM) has been implemented for the signal peptide (SP) prediction task in eukaryotes. The implementation was carried out paying attention to one of the main difficulties that characterizes this task, the presence of other N-terminal hydrophobic structures that can be mislabeled as SP. To this end, the SVM is implemented using and combining several features (i.e 'hp', 'ghp','tmt') that allow the analysis of the hydrophobic character of each window. Moreover, the feature extraction has been implemented on the entire length, to get the broader context of the protein. These arrangements resulted in a model that overperformed when compared to a position-specific weight matrix (PSWM) method. Indeed, with respect to the latter, the SVM improved the MCC of almost 0.20 and significantly reduced the FPR by 4%.

Even though the reduction of FPR_{TM} with respect to the von Heijne method is remarkable, the presence of TM mislabeled as SP indicates that this method comes with limitations when dealing with transmembrane regions at the N-terminus and that more powerful and complex approaches are needed. Indeed, to take the level of prediction further, deep neural networks came into play in the last few years. Their many layers can filter and reorder features in very powerful ways (Savojardo et al., 2018; Almagro Armenteros et al., 2019) and can recognize also 'non-standard' SPs. Moreover, recently SignalP 6.0 has been implemented through the utilization of protein language models (pLMs), which allowed the prediction of all five types of SPs with a limited amount of training data (Teufel et al., 2022).

Nonetheless, even with a simpler approach than those just mentioned, the SVM model presented here can still be regarded as achieving high performance, precision, and sensitivity for the task it was designed for.

References

Almagro Armenteros,J.J. et al. (2019) SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat. Biotechnol., 37, 420–423.

Boser,B.E. et al. (1992) A training algorithm for optimal margin classifiers. In, Proceedings of the fifth annual workshop on Computational learning theory, COLT '92. Association for Computing Machinery, New York, NY, USA, pp. 144–152.

Choo,K.H. and Ranganathan,S. (2008) Flanking signal and mature peptide residues influence signal peptide cleavage. BMC Bioinformatics, 9, S15.

Chou,P.Y. and Fasman,G.D. (1978) Prediction of the secondary structure of proteins from their amino acid sequence. Adv. Enzymol. Relat. Areas Mol. Biol., 47, 45–148.

Crooks,G.E. et al. (2004) WebLogo: a sequence logo generator. Genome Res., 14, 1188–1190.

Dalbey,R.E. et al. (2012) Membrane Proteases in the Bacterial Protein Secretion and Quality Control Pathway. Microbiol. Mol. Biol. Rev. MMBR, 76, 311–330.

Dalbey,R.E. et al. (1997) The chemistry and enzymology of the type I signal peptidases. Protein Sci., 6, 1129–1138.

von Heijne,G. (1986) A new method for predicting signal sequence cleavage sites. Nucleic Acids Res., 14, 4683–4690.

Kovjazin,R. et al. (2011) Signal peptides and trans-membrane regions are broadly immunogenic and have high CD8+ T cell epitope densities: Implications for vaccine development. Mol. Immunol., 48, 1009–1018.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. J. Mol. Biol., 157, 105–132.

Nielsen,H. et al. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. Des. Sel., 10, 1–6.

Nielsen,H. and Krogh,A. Prediction of Signal Peptides and Signal Anchors by a Hidden Markov model.

Nilsson,I. et al. (1994) The COOH-terminal ends of internal signal and signal-anchor sequences are positioned differently in the ER translocase. J. Cell Biol., 126, 1127–1132.

Ohmuro-Matsuyama,Y. and Yamaji,H. (2018) Modifications of a signal sequence for antibody secretion from insect cells. Cytotechnology, 70, 891–898.

Perlman,D. and Halvorson,H.O. (1983) A putative signal peptidase recognition site and sequence in eukaryotic and prokaryotic signal peptides. J. Mol. Biol., 167, 391–409.

Robinson,P.J. and Bulleid,N.J. (2020) Mechanisms of Disulfide Bond Formation in Nascent Polypeptides Entering the Secretory Pathway. Cells, 9, 1994.

Savojardo,C. et al. (2018) DeepSig: deep learning improves signal peptide detection in proteins. *Bioinformatics*, **34**, 1690–1696.

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. Nucleic Acids Res., 18, 6097–6100.

Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol., 35, 1026–1028.

Teufel,F. et al. (2022) SignalP 6.0 predicts all five types of signal peptides using protein language models. Nat. Biotechnol., 40, 1023–1025.

The UniProt Consortium (2023) UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res., 51, D523–D531.

Vert,J.P. (2002) Support vector machine prediction of signal peptide cleavage site using a new class of kernels for strings. Pac. Symp. Biocomput. Pac. Symp. Biocomput., 649–660.

Von Heijne,G. (1983) Patterns of Amino Acids near Signal-Sequence Cleavage Sites. Eur. J. Biochem., 133, 17–21.

Von Heijne,G. (1990) The signal peptide. J. Membr. Biol., 115, 195–201.

Zhao,G. and London,E. (2006) An amino acid “transmembrane tendency” scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: Relationship to biological hydrophobicity. Protein Sci. Publ. Protein Soc., 15, 1987–2001.