Sequence analysis

Signal peptide prediction in eukaryotes: a comparison of SVM with a PSWM-based method

Mario Esposito[1,*]

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

*To whom correspondence should be addressed.

Abstract

Motivation: Signal peptides (SP) are short peptides located at the N-terminus of proteins, carrying information for protein secretion. SPs gained attention in several scientific and industrial applications, including recombinant protein production and immunization. Therefore, since the dawn of bioinformatics, the prediction of signal peptides and protein subcellular location has been one of the main focal points. In this work two different sequence-based methods for eukaryotes SP proteins classification were implemented. A support vector machine (SVM) classifier trained with a combination of features describing residue composition and local hydrophobicity. The position-specific weight matrix (PSWM) method based on von Heijne (VH) method was implemented for comparison.

Results: VH method achieved a Matthew's correlation coefficient of 0.68, while SVM method outperformed VH with an MCC of 0.89 and lower false positive rate (FPR = 0.9%).

Contact: mario.esposito17@studio.unibo.it

1 Introduction

The best-known protein 'zip code' is the secretory signal peptide (SP), which is found in all the three domains of life. It targets a protein for translocation across the plasma membrane in prokaryotes and across the endoplasmic reticulum membrane in eukaryotes (von Heijne, 1990; Emanuelsson et al., 2007). It is an N-terminal peptide, typically 15–30 residues long, which is cleaved during translocation of the protein across the membrane. It is important to emphasize that the presence of an SP does not necessarily means that the protein is secreted, but it only means that it enters the secretory pathway. High variation in SP composition is responsible for their high capacity for protein translocation. There is no simple consensus sequence for SP, but they typically show three distinct regions: an N-terminal region (N-region), a hydrophobic region (H-region) and a C-terminal region (C-region)(Figure 1). The N-region highly varies in length but usually it consists of ~1-5 residues and extensive analysis of bacterial and eukaryotic SPs depicted a greater inclination for a net positive charge within prokaryotic N-regions, and a lower one in eukaryotic SPs (Choo and Ranganathan, 2008). The H-region forming the hydrophobic core at the center of the SP is lined with stretches of ~7-15 hydrophobic residues (mainly leucine) and it tends to adopt an α-helical conformation (Nielsen et al., 1997). The C-terminal part of an SP is called C-region consisting of ~3–7 neutral or polar residues. The critical points of the C-region are positions -1 and -3 prior to the cleavage site (CS), well-recognized as the (−3, −1) rule or AXA|VXA motif (von Heijne, 1984). This motif is universal in prokaryotic, ER, and organellar SPs, however, its conservation in eukaryotes is lower than in prokaryotes (von Heijne, 1990; Owji et al., 2018). SP prediction involves two sub-tasks: discriminating between SPs and non-SP proteins and predicting the position of the SP cleavage site. A major challenge in signal peptide detection lies in the differentiation

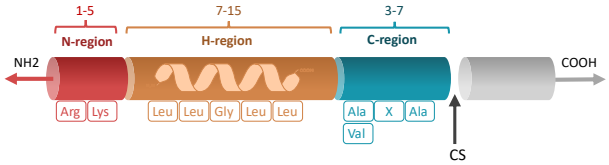

Figure 1. SP structure depicting the three characteristic regions and their typical length and residue composition. The cleavage site (CS) is depicted between the C-region and mature protein (grey).

between authentic SP H-region and N-terminal transmembrane helices since both regions present a similar residue composition and hydrophobicity propensity. Also, chloroplasts and mitochondria signal peptides known as transit peptides share some similarities with the SP structure. Moreover, accurately detecting the cleavage site poses another challenge, primarily due to the signal sequence's considerable length variation and the absence of unequivocal sequence motifs pinpointing the cleavage site's location. When constructing any prediction method, a key factor is the quality of the data, and the training set construction from available databases implies a large amount of work and carries a number of critical decisions. So, another challenge is for example a critical decision of adding or not adding a selection criterion, which often represents a trade-off between a more conservative approach and preserving more training samples. Since the dawn of bioinformatics research, the prediction of signal peptides and protein subcellular location from amino acid sequences has caught significant interest, leading to the evolution of numerous statistical and machine learning methodologies. The first prediction method was described by von Heijne in the paper that introduced the (−3, −1) rule in 1983 (Von Heijne, 1983). Some years later, the same author proposed a PSWM

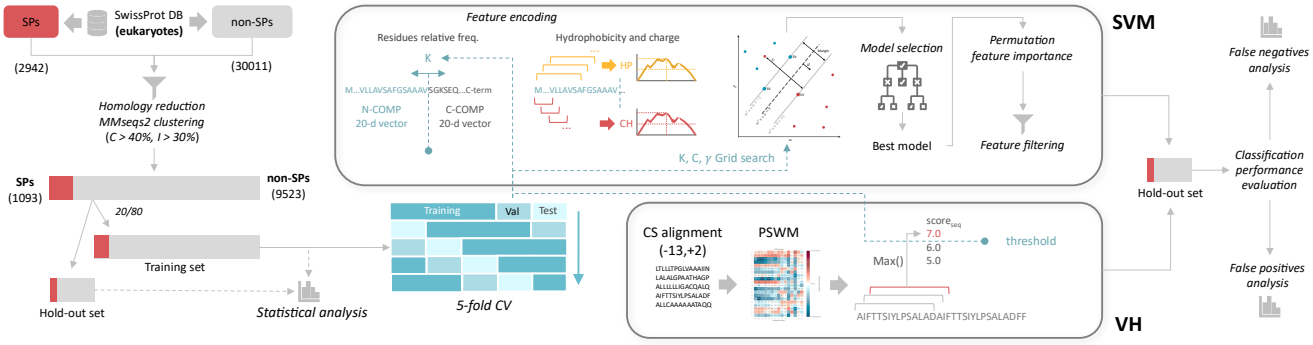

Figure 2. Workflow of the study.

based method to detect the SP CS (von Heijne, 1986). SignalP 1.0 was in 1996 one of the first implementations of artificial neural networks (ANNs) for SP prediction adopting two different ANNs with moving windows (Nielsen et al., 1997). SignalP 2.0 and 3.0 adopt in addition to ANNs, an hidden Markov model (HMM) module to discriminate between cleaved SP from uncleaved signal anchors (Nielsen and Krogh, 1998; Dyrløv Bendtsen et al., 2004). Although support vector machines (SVMs) had not proven to be as effective as ANNs, several attempts were made, having as first relevant SVM implementation in 2002 (Vert, 2002). It is interesting to note also MEMSAT-SVM, which proved to be better than its predecessor based on ANNs (Nugent and Jones, 2009). Also the combination of SVM and ANNs was proposed, adopting SVM as a primary classifier to discriminate SP from non-SP sequences and employing NNs to predict the most suitable cleavage site (Kazemian et al., 2014). Deep learning in SP prediction was introduced by DeepSig (Savojardo et al., 2018), based on convolutional ANNs and followed by SignalP 5.0 based on recurrent ANNs (Almagro Armenteros et al., 2019). Recently the last version of SignalP (6.0) adopted protein language models to predict all five types of signal peptides (Teufel et al., 2022).

This study aimed to implement two different sequence-based methods to classify signal peptide (SP) proteins in eukaryotes. The reference VH method adopted a position-specific weight matrix (PSWM) to model the cleavage site (CS) derived from a multiple sequence alignment of SP fragments, and a sliding window to score and classify sequences. The proposed SVM was trained with a combination of features, including N-terminal residue composition, global residue composition and local hydrophobicity, to classify proteins as SP or non-SP. Both methods' hyperparameters were optimized through a 5-fold CV and the classification performances were benchmarked adopting a hold-out test. The overall process is displayed in Figure 2. VH Method achieved a Matthew's correlation coefficient (MCC) of 0.68 while SVM method outperformed VH with an MCC of 0.89, a lower false positive rate (FPR) and higher precision and recall.

2 Methods

2.1 Datasets

2.1.1 Download data

Data were downloaded from UniProtKB/SwissProt release 2023_04 (The UniProt Consortium, 2023), including only proteins longer than 30 residues in Eukaryotes. The positive set (SP proteins) was constructed by selecting only proteins with an experimentally annotated SP cleavage site at positions greater or equal than 13. The negative set (non-SP proteins) was constructed selecting proteins without SP (at any evidence level) and annotated at experimental evidence as belonging to cellular compartments unrelated to SP: cytosol, nucleus, mitochondrion, plastid, peroxisome and cell membrane. In addition, all proteins whose annotations contained words related to the secretory pathway were excluded (endoplasmic | Golgi | secreted | lysosome). In the end the positive and negative sets included 2942 and 30011 proteins, respectively.

2.1.2 Datasets generation

All sequences were homology-reduced to obtain non-redundant datasets by applying a clustering procedure with MMSeqs2 (Steinegger and Söding, 2017), with a minimum identity threshold of 30% and a minimum pairwise coverage threshold of 40%. A connected component clustering mode was chosen, so that if two proteins are clustered with a third one, they both end up in the same set. In the end only representatives for each cluster were retained. To assess the generalization performance of the methods, 20% of the data was retained as a hold-out set (benchmarking) and 80% of the data was adopted for the training procedures. Furthermore, to perform the cross-validation (CV) procedures the training set was divided in 5 subsets. Both splitting procedures were applied randomly and maintaining the same ratio of SP and non-SP proteins. The length distributions of SPs were plotted for each of the 5 training subsets to examine the homogeneity of partitioning (Supplementary Figure S1).

Table 1. Dataset composition.

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

2.1.3 Transmembrane and transit peptide annotation

To carry out the false positive analysis after benchmarking, non-SP proteins were annotated as transmembrane (Tm) and/or transit peptide (Ts) based on UniProtKB/SwissProt annotation. Only manually curated or experimental annotations were considered (ECO:0000269, ECO:0000303, ECO:0000305, ECO:0000250, ECO:0000255, ECO:0000312, ECO:0007744). A protein was considered as Tm if at least one transmembrane region was located in the first 90 residues of the sequence. A protein was considered as having a Ts if it was present in the first 90 residues.

2.1.4 Statistical analysis

To test whether random splitting of the data into the benchmarking and training set had generated homogeneous distributions, several features were compared between training and benchmarking sets. In addition, to detect specific characteristics of SP fragments, some features were

compared between SP and non-SP proteins. No obvious difference between training and benchmarking distributions was detected at any feature level (Figure 3). This showed that the benchmarking set is a good representation of the whole dataset and they could be adopted for performance evaluation. Considering the SP length distribution, the maximum, minimum and average SP length are 64, 13, and ~23, respectively (Figure 3A). CS sequence logos for SP proteins were produced separately for both benchmarking and training with WebLogo (Crooks *et al.*, 2004; Schneider and Stephens, 1990), taking into account residues from position -13 to +2 relative to the CS position (Figure 3C, 3D). The two CS logos look very similar to each other, demonstrating an homogenous split, furthermore, both are in agreement with previous knowledge, showing conservation of the H-region (-13, -6 and the AXA|VXA motif (-3, -1), although this motif is more conserved in prokaryotes (Von Heijne, 1983; Perlman and Halvorson, 1983; Choo and Ranganathan, 2008). The sequence length distributions were compared between SP and non-SP proteins (Figure 3B), showing an average slightly longer sequences for non-SP, which does not seem relevant.

The residue composition of the SP fragments was compared with a background distribution computed considering the residue composition of all the eukaryotic sequences in UniProtKB/SwissProt (Figure 3E). Considering the difference in residue composition in respect to the background distribution , leucine, alanine, valine and methionine are clearly overrepresented in SP fragments. Methionine overrepresentation is biased, because SP fragments are located at the N-terminus of the sequence where methionine is almost always present as sequence starting residue. On the other hand, it is relevant that the hydrophobic residues (L, A and V) are overrepresented consistently with the presence of the SP H-region. It is worth to notice that charged residues and some polar residues seem to be underrepresented, which could be naturally a consequence of the overrepresentation of the hydrophobic residues.

In addition, to understand if there could be some relation between the presence of the SP and the global residue composition, the latter was compared between SP and non-SP proteins (Supplementary Figure S2). Although it is complex to interpret, differences can be noticed in the percentage of some residues (G, E, K, R, S and C). A relevant overrepresentation of cysteine (1.65% in non-SP vs 2.85% in SP) could be explained considering that disulfide bonds are mostly confined in proteins belonging to the secretory pathway and SP is strongly related with this pathway (Bošnjak *et al.*, 2014; Robinson and Bulleid, 2020). So, global residue composition can be adopted as a feature in the SVM classification method.

In the whole dataset the most represented kingdom and species are respectively Metazoa and *Homo Sapiens* (Supplementary Figure S3). In addition, the distributions of the kingdoms and top 10 species were also compared between SP and non-SP proteins, and it is possible to notice that SP protein showed a skewed kingdom distribution towards Metazoa (Supplementary Figure S4B) and a skewed species distribution towards *Homo Sapiens* (Supplementary Figure S4A). This could be due to the fact that in general SP is more studied (thus more annotated) in humans and Metazoa.

2.2 The von Heijne (VH) method

Weight-matrix methods have been used for a number of years to locate signals in protein sequences, and regarding the SP detection the original weight-matrix method was proposed by von Heijne in 1983 (von Heijne, 1986). Basically, the idea is to collect a multiple sequence alignment of the SP fragments comprising the CS from position -13 to position +2 and compute for each position of the alignment, the relative frequency of each residue, obtaining a position-specific probability matrix (PSPM) (1). After that, the PSPM has to be normalized against a background frequency distribution, becoming a position-specific weight matrix (PSWM) (2). Once the PSWM is modeled, it can be adopted to scan over a query sequence computing a score for each position of the sliding window (3). The highest score position will be the putative CS localization and the highest score will be assigned to the protein. In the end if this score is greater than a threshold the query protein can be classified as SP, otherwise it can be classified as non-SP. The following method is based on the original method with a few modifications, and it was adopted only for SP classification and not for labeling.

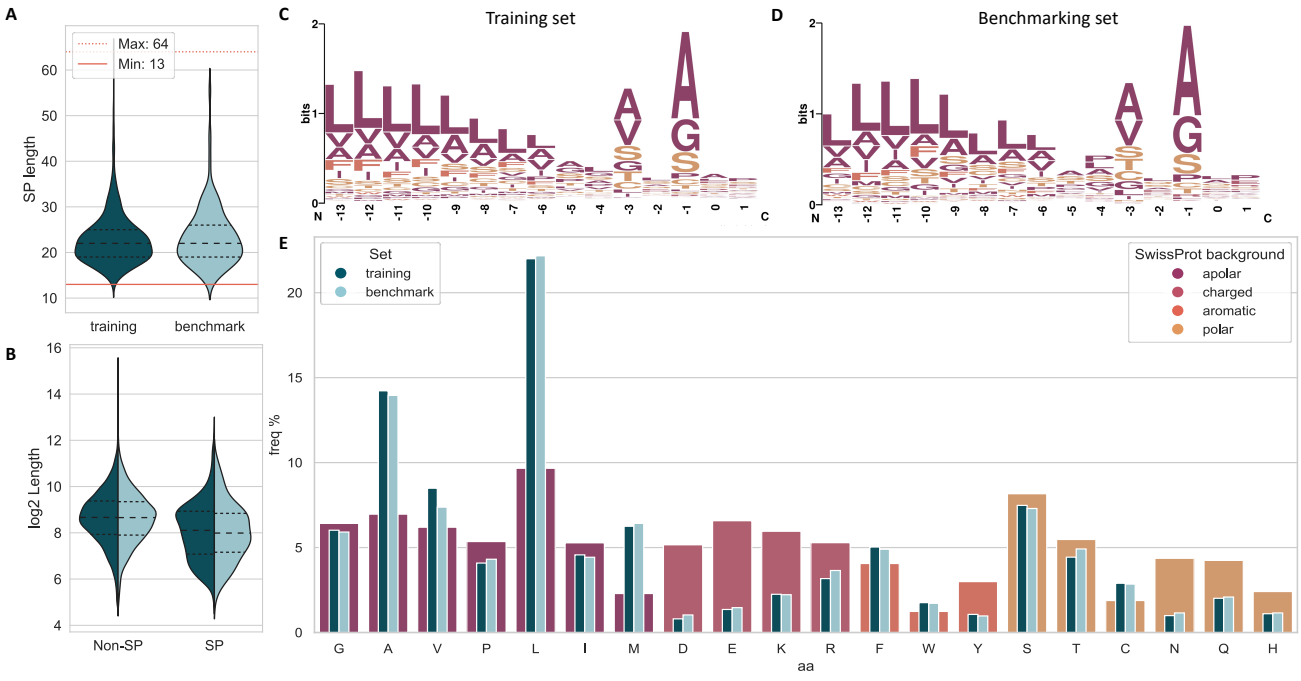

Figure 3. A. SP length distribution (training vs benchmark). **B.** Sequence (log_2) length distribution by class and set **C.** SP cleavage site sequence logo (-13,+2) computed within the training set **D.** SP cleavage site sequence logo (-13,+2) computed within the benchmarking set **E.** The residue composition of the SP fragments (training vs benchmarking) compared with a background distribution computed considering the residue composition of all the eukaryotic sequences in UniProtKB/SwissProt (on x-axis residue letters are grouped by similar physico-chemical properties).

2.2.1 Describing the algo

Given a set of N aligned sequences of length 15, the PSPM M is computed 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 1 if the condition is met, 0 otherwise

Pseudocounts are added to avoid 0 values which could cause problems in the subsequent logarithm computation (2). From the PSPM M (1), the PSWM W is computed as follows:

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

Where b_k is the relative frequency of residue k in the background model (overall SwissProt eukaryotes sequences). To score a new sequence, for each subsequence X = (X_1,…,X_L) of length 15 obtained by scanning a sliding window from position 1 to 90-15, the (log-likelihood) score of X given the PSWM W is computed as follows:

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

Among all the subsequences $score_{(X|W)}$, only the highest one will be assigned to the sequence and if this score is greater than or equal to the threshold, the sequence is classified as SP, otherwise it is classified as non-SP.

2.2.2 Threshold optimization (CV), training and benchmarking

The splitting procedure is described in section 2.1.2. In order to optimize the threshold hyperparameter, a 5-fold CV was performed over the training set by the following strategy. For each CV run, 3 subsets of the training set were adopted to train the model, one subset was adopted to select the score threshold maximizing the F1 score (validation subset) and another subset was adopted to test the performance adopting the selected score threshold (testing subset). The precision-recall curve and ROC curve of the CV procedure are in Supplementary Figure S5.

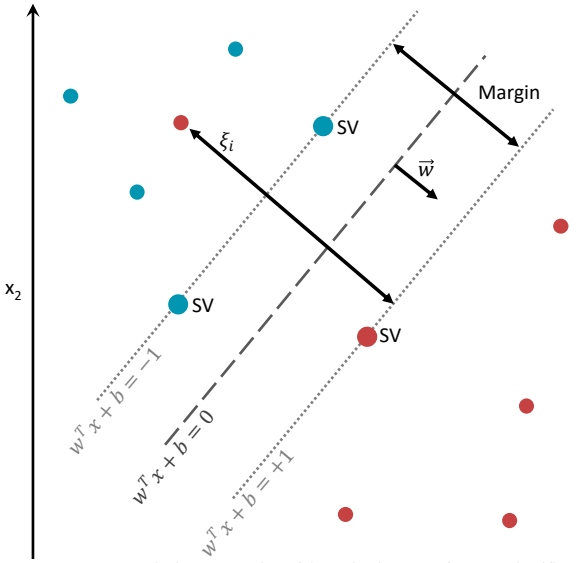

Figure 4. Geometrical representation of the main elements of a SVM classifier in 2D. SVs are the support vectors, corresponding to the data sample whose α is positive.

The model was trained computing the PSWM over the whole training set (Supplementary Figure S6) and the average optimal threshold among CV runs was selected to evaluate the classification performance on the benchmarking set.

2.3 SVM

2.3.1 Support Vector Machine for classification with soft margins

Support vector machines (SVMs) (Boser et al., 1992) are powerful classification algorithms that have shown valuable performance in a variety of biological classification tasks, such as gene expression analysis, protein classification, and disease diagnosis. The objective of the SVM algorithm is to find the hyperplane (4) (decision boundary) maximizing the margin in an N-dimensional space that distinctly classifies the data points (5). With soft margin formulation (6), SVM allows for some misclassifications, keeping margins as wide as possible so that other points can still be classified correctly, allowing SVMs to generalize well to unseen data. To make SVM handle non-linear data, the kernel trick is adopted (7,9). It consists in implicitly mapping data into a higher-dimensional space, enabling the discovery of non-linear decision boundaries. This is crucial for dealing with complex datasets where simple linear separation is not possible.
The separating hyperplane is represented by the equation:

$$(w^T x + b) = 0 \quad (4)$$

Where:

- w is the weight vector perpendicular to the hyperplane
- b is the bias term
- x is the input vector

SVM imposes constraints to ensure the margin between classes:

$$y_i\,(w^T x_i + b) \geq 1 - \xi_i \quad (5)$$

Where:

- y_i is the class label,
- x_i is a data point, and
- ξ_i are non-negative slack variables.

The objective function to minimize for soft-margin SVM is:

$$\min_{w\,b\,\xi} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \quad (6)$$

Where C is the regularization hyperparameter.
The optimization problem involves the maximization of the dual Lagrangian:

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

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

Where:

- α is a vector of Lagrange multipliers, and
- $K(x_i, x_j)$ is the kernel function

For this method the selected function for the kernel is the gaussian radial basis function (RBF):

$$K(x_i, x_j) = exp\left(-\gamma \left\| x_i - x_j \right\|^2\right), \quad \gamma = \frac{1}{2\sigma^2} \qquad (9)$$

Where:

- γ is a hyperparameter that determines the scale of the kernel
- $\| x_i - x_j \|^2$ is the squared euclidean distance.

The RBF kernel is a flexible choice in Support Vector Machines (SVM) for capturing non-linear relationships in data measuring the similarity between data points by computing the exponential of the negative squared euclidean distance.

2.3.2 Input encoding and features

The SVM application in the context of SP proteins classification consisted of encoding input sequences in n-dimensional vectors and assigning 0 and 1 as class label to Non-SP and SP respectively. Each input sequence was encoded in 5 different feature groups:

- N-terminal residue composition (N-COMP) encoded as a 20-dimensional vector where each component represents a different residue relative frequency computed taking the region of the input sequence from position 1 to K.
- C-terminal residue composition (C-COMP) encoded as a 20-dimensional vector where each component represents a different residue relative frequency computed taking the region of the input sequence from position K +1 to the end.
- Local hydrophobicity type 1 (HP) is a propensity feature encoded as a 3-dimensional vector where each component represents average, maximum and maximum position of the profile produced by the Kyle and Dolittle hydropathy scale (Kyte and Doolittle, 1982). The profile was computed by scanning the sequence from position 1 to K adopting a sliding window of length 7 (without padding).
- Local hydrophobicity type 2 (HP2) is a propensity feature encoded as a 3-dimensional vector where each component represents average, maximum and maximum position of the profile produced by the Kyle and Dolittle hydropathy scale. The profile was computed by scanning the sequence from position 1 to 40 adopting a sliding window of length 12 (without padding).
- Local positive charge (CH) is a propensity feature encoded as a 3-dimensional vector where each component represents average, maximum and maximum position of the profile produced by a trivial positive charge scale (His=1, Arg=1, Lys=1, all the other residues = 0). The profile was computed by scanning the sequence from position 1 to K adopting a sliding window of length 5 (without padding).

The Kyte and Doolittle hydropathy scale values were normalized between 0 and 1, to be consistent with the other features. The HP2 feature was introduced to refine HP in detecting the H-region because some SP were much longer than the average, thus instead of stopping at position K, the sliding window stopped at 40 for HP2. In addition, it worked better with a wider sliding window (L = 12).

2.3.3 Hyperparameters and model selection

The hyperparameter γ controls the kernel's flexibility: smaller γ results in a smoother, more flexible decision boundary, while larger γ creates a more localized, less flexible boundary. Hyperparameter C in an SVM involves finding the balance between maximizing the margin and minimizing classification errors. The hyperparameter K adopted in the feature encoding basically reflects the average CS position (~23). An exhaustive grid-search 5-fold CV was performed for each different model in order to select the optimal hyperparameters combination. The 3 hyperparameters lists (Supplementary Table S2) were constructed in order to avoid that optimal results would fall at the extremes of the range and CV procedure was applied with the same subsets and criteria of VH method (paragraph 2.2.2). In particular, for each run, the SVM was trained on 3 subsets with all 330 hyperparameter combinations and the combination maximizing the MCC on the validation subset was taken as the best combination. The best combination was adopted to retrain the model on the 3 subsets and test it on the testing subset. For each model the hyperparameters combination will be the most frequent value of each hyperparameter among the runs. In case of tie, for C and γ the lowest value is picked, whereas for K the value that is closest to the average length of SP (23).

To find the optimal features combination (model selection), the following heuristic strategy was adopted:

1. N-COMP is taken as only feature for the base model and a grid-search CV is performed to optimize the hyperparameters and the average MCC is assigned to the base model.
2. For each remaining feature in the set.
   a. add the feature to the base model.
   b. perform the grid-search CV to optimize the hyperparameters and compute the average MCC.
   c. compute the difference between the average MCC and the base model (ΔMCC).
3. Feature with highest ΔMCC is added to the base model and removed from the set.
4. Features with a negative ΔMCC are discarded from the set.
5. Iterates from step 2 starting with new base-model, until no features have a positive ΔMCC or the set is empty.

2.3.4 Permutation feature importance

Once the best model was selected, to detect the most important features and eventually filter out the irrelevant ones, a permutation feature importance (PFI) was carried out. PFI is defined to be the decrease in a model score when a single feature value is randomly shuffled (Breiman, 2001). This procedure breaks the relationship between the feature and the target, thus the drop in the model score is indicative of how much the model depends on the feature. This technique benefits from being model agnostic and can be calculated many times with different permutations of the feature. In particular, MCC was adopted as model score and each feature was randomly permuted 20 times. PFI was performed in CV adopting 3 subsets for the training and 1 subset for the PFI. Each CV run produced a partially overlapping feature ranking , so, in order to obtain a consensus, the average rank was computed for each feature (Supplementary Table S3). To filter out the less important features only the top-k features of the average ranking were retained. To optimize k, another CV was performed training the model only with top-k features and computing the MCC on the testing subset. In the end the smallest k which maintained invariant the MCC in respect to no feature filtering was selected (Supplementary Figure S7).

2.4 Performance metrics

TP, TN, FP and FN are true positive, true negative, false positive and false negative predictions, respectively.

In order to assess the performances, the following metrics were used:

- Accuracy:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad (10)$$

- Matthews correlation coefficient (MCC):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (11)$$

- Precision:

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

- Recall (TPR):

$$Recall = TPR = \frac{TP}{TP + FN} \quad (13)$$

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

$$F1\ score = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (14)$$

- FPR (false positive rate):

$$FPR_X = \frac{TP_X}{TP_X + FP_X} \quad (15)$$

Where X is the set in which the predictions are considered and it can be equal to: the whole benchmark, the transmembrane proteins (Tm) or the transit peptide proteins (Ts).

- FNR (false negative rate):

$$FNR = \frac{FN}{TP + FN} = 1 - TPR \quad (16)$$

3 Results

3.1 CV results

A 5-fold CV procedure applied to optimize the score threshold in the VH method (see section 2.2.2), resulted in an average MCC of 0.69 ± 0.01 (average ± std. error)(Table 2). In addition, for each SVM model, a grid search 5-fold CV was performed in order to optimize hyperparameters K, γ and C (see section 2.3.3) for each model. The baseline SVM, while having as features only the N-COMP, outperforms the VH method giving a MCC of 0.79 ± 0.01. Furthermore, the MCC significantly improves when combining other feature groups (see section 2.3.2 and 2.3.3), indeed, adding the HP2 alone, increases the MCC to 0.87 ± 0.01, supporting the importance of the H-region in the SP's structure. It is worth noting how the addition of HP2 to the base model resulted in a higher performance, with respect to the addition of the analogous HP feature (MCC = 0.84 ± 0.01). This result supports the better designing of HP2 which has a wider sliding window (SW)(L = 12) than HP (L = 7) and a longer scanning of the SW (until position 40 rather than K = 27). The rationale behind this improvement could be that a SW which scans until 40 characterizes better the subset of SPs longer than 27 (FNR 3% lower) and a wider SW can average better the hydrophobic profile, improving the detection of H-regions. On the other hand, the addition of the CH feature does not result in a significant improvement (MCC = 0.80 ± 0.02), suggesting that a better featuring of the N-region could be implemented.

Table 2. CV results for all the models (model name in bold, highlight the baseline SVM to which features were added).

Model	MCC test	Precision	Recall
SVM (N-COMP)	**0.79±0.01**	0.84±0.01	0.79±0.01
SVM+C-COMP	0.83±0.01	0.88±0.01	0.83± 0.01
SVM+HP	0.84 ± 0.01	0.86 ± 0.01	0.86 ± 0.01
SVM+HP2	**0.87 ± 0.01**	0.87 ± 0.00	0.89 ± 0.01
SVM+CH	0.80 ± 0.02	0.86 ± 0.01	0.79 ± 0.03
SVM+HP2+C-COMP	**0.89 ± 0.00**	0.91 ± 0.01	0.90 ± 0.01
SVM+HP2+HP	0.88 ± 0.01	0.88 ± 0.01	0.90 ± 0.01
SVM+HP2+CH	0.87 ± 0.01	0.88 ± 0.01	0.88 ± 0.02
SVM+HP2+C-COMP+HP	0.89 ± 0.01	0.89 ± 0.01	0.91 ± 0.01
SVM+HP2+C-COMP	**0.89 ± 0.00**	0.91 ± 0.01	0.90 ± 0.01
SVM+HP2+C-COMP+FF	**0.89 ± 0.01**	0.90 ± 0.01	0.91 ± 0.02
VH	**0.69 ± 0.01**	0.73 ± 0.03	0.72 ± 0.03

(The extended table with accuracy, MCC on validation and best hyperparameters is in Supplementary Table S4)

By taking the SVM model N-COMP+HP2 as the new baseline model, and adding the C-COMP feature, it improves more than the other feature groups (MCC = 0.89 ± 0.00). This result confirms the hypothesis that there could be some relation between the presence of the SP and global residue composition (see section 2.1.4). In the end, N-COMP+HP2+C-COMP is considered as the best feature combination, since the addition of HP does not yield better performance.

3.2 Permutation feature importance and feature filtering

To detect the most important features (among 43) and eventually filter out irrelevant ones, a permutation feature importance (PFI) was carried out (see section 2.3.4). Five partially overlapping feature importance rankings produced by PFI in CV were averaged into one, computing the average rank for each feature (Supplementary Table S3). To design a parsimonious feature filtering, the top 28 features were retained for the feature filtering, maintaining the same MCC of the original model. In the end, the new model with feature filtering (N-COMP+HP2+C-COMP+FF) was selected as the final model to perform the benchmarking on the hold-out set (Table 2). Maximum hydrophobicity (HP2), maximum hydrophobicity position (HP2), leucine frequency (N-COMP) and average hydrophobicity, obtain an average rank respectively of 1.0, 2.0, 3.2 and 4.0, highlighting the importance of HP2 in discriminating SP from non-SP. Overall, 15 residue frequencies (features) incorporated into the N-COMP feature group rank in the top-28 features by the following order: L, R, W, K, I, I, D, S, E, G, Q, N, V, P, Y, H. Considering C-COMP, 10 residue frequencies rank in the top-28 features by the following order: K, I, C, M, R, L, G, Y, N, D (Supplementary Table S3). One of 10 residues was cysteine, which was shown to be globally overrepresented in SP proteins (1.65% in non-SP vs 2.85% in SP)(Supplementary Figure S2). This observation is coherent with the fact that disulfide bonds are mostly confined in proteins belonging to the secretory pathway which is related to SP (Bošnjak et al., 2014; Robinson and Bulleid, 2020).

3.3 Benchmarking

The final VH model was trained computing the PSWM over the whole training set (80% of the dataset) (Supplementary Figure S6) and the average optimal threshold was selected to evaluate the classification performance on the hold-out set (20%), yielding an MCC of 0.68 (Table 3).
The final SVM model (N-COMP+HP2+C-COMP+FF) selected in CV was trained with the entire training set adopting the feature filtering and the following optimal hyperparameters: K=27, C=8 and γ=8.
When SVM is tested on the benchmarking set, it performs with an MCC of 0.89. SVM outperforms the VH method as in CV, with improvement in precision and recall of 21% and 17%, respectively. Both methods show almost no change in performance when compared to their correspondent CV, demonstrating a good generalization capacity and no overfitting symptoms. Furthermore, the feature filtering applied to the SVM does not seem to affect performance in benchmarking.

Table 3. Final models classification performances in benchmarking.

Model	MCC	Accuracy	Precision	Recall
VH	0.68	0.94	0.71	0.72
SVM+HP2+C-COMP+FF	0.89	0.98	0.92	0.89

(Confusion matrices are in Supplementary Figure S8)

3.4 False positives and false negatives analysis

3.4.1 False positives analysis

The VH method yields a FPR of 3.4% while SVM shows a lower FPR of 0.9% (Table 4). Since the H-region of the SP can be confused with transit peptide (Ts) and transmembrane (Tm) regions, it can lead to misclassifications. In order to investigate it, non-SP proteins in the benchmarking set were annotated as being Tm (n = 160) and/or Ts (n = 218) and the FPRs within Tm and Ts subset were computed (see section 2.1.3).
The VH method shows a high FPR_{Tm} of 18.8%, highlighting PSWM limitation in discriminating between the Tm region and SP H-region (Supplementary Figure S9A). In addition, comparing the proportions of Tm proteins between TN and FP, the FP set shows a significant enrichment in Tm proteins (Fisher exact p = 5.5e-17, OR = 11.6)(Supplementary Figure S9B). The SVM method shows a FPR_{Tm} of 6.3% showing an improvement compared to VH (Supplementary Figure S9D). However, comparing the proportions of Tm proteins between TN and FP, the FP set shows a significant enrichment in Tm proteins (Fisher exact p = 3.2e-7, OR = 14.5)(Supplementary Figure S9E).
On the other hand, the FPR_{Ts} is 4.6% and 0.5% in VH and SVM respectively, and both values are very similar to the corresponding global FPR. Indeed, comparing the proportions of Ts proteins between TN and FP, no significant difference in Ts proportion is found for both methods (Supplementary Figure S9C, F), suggesting that Ts regions were not a main reason for SP misclassification.

Table 4. Final models classification performances on the hold-out set.

Model	FPR	FPR_{Tm}	FPR_{Ts}
VH	3.4%	18.8% (30)	4.6% (10)
SVM+HP2+C-COMP+FF	0.9%	6.3% (10)	0.5% (1)

3.4.2 False negatives analysis

In terms of recall (TPR), the SVM method improves by 17% compared to VH with a respective FNR of 11% and 28% (FNR = 1-TPR). To highlight factors that potentially affect the detection of SP proteins by the VH method, the sequence logo was computed only taking the FN cleavage site (CS) (-13,+2) sequences (Supplementary Figure S10 B). The FN CS sequence logo shows less information content, when compared to the general CS sequence logo (Supplementary Figure S10 A), in both in the H-region (-13, -6) and in the AXA|VXA motif (-1,-3).
Regarding potential factors affecting detection performance in the SVM, both SP length heterogeneity and N-terminal residue composition (until position K = 28) are examined. The SP length (CS position) distributions are compared between FN and TP (Supplementary Figure S11 A). Both distributions do not fit a gaussian distribution (Shapiro's $p < 0.05$); a significantly higher variance is detected in FN (Levene's p = 1e-6) and a different shape distribution is also detected (MWU p = 0.014). Therefore, since the FN set also shows a higher median than TP, these results suggest that the hyperparameter K behaves like constraints in defining the CS position, penalizing the detection capability of longer SP. N-terminal residue compositions (until position K = 28) are compared between FN and TP (Supplementary Figure S11 B), showing no overlap between some residues in terms of relative frequencies (G, A, L, D, E, K, R, N, H). This observation suggests that the SP composition heterogeneity within SP sequences could be a reason for missing some SPs as TP.

4 Conclusion

In summary, in this work two different sequence-based methods were implemented to classify SP proteins in eukaryotes. Both VH and SVM hyperparameters were optimized by a 5-fold CV on the training set (80%) and benchmarked on the hold-out set (20%). The proposed SVM classifier was trained with a combination of features describing both N-terminal/C-terminal (global) residue composition and local hydrophobicity. Furthermore, a PFI was performed to detect the most important features and implement a feature filtering. Comparing the SVM to the reference PSWM-based method (VH), VH achieved a MCC = 0.68, while the SVM outperforms it with an MCC of 0.89. The better SVM performances could be explained by the fact that the VH was based only on CS properties, while the SVM was trained with several features. Overall, both methods results do not show noticeable difference between CV and benchmarking performances, demonstrating the efficacy of all the practices applied to avoid overfitting. Although SVM shows an improvement in FPR (0.9%) when compared with VH (FPR = 3.4%), transmembrane proteins seem to remain a key factor in SP misclassification. Also, SVM recall (FNR = 11%) shows a relevant improvement compared to VH (FNR = 28%). However, SP length and residue composition heterogeneity still seems to remain a problem in classifying some SPs as positives. Therefore, in order to improve both recall and precision capability it would be appropriate to switch to other approaches, such as ANN, or even more sophisticated approaches as deep learning and protein language models (pLMs). Deep learning, characterized by neural networks with multiple layers, offers a powerful framework for intricate pattern recognition. For instance, convolutional ANNs in DeepSig (Savojardo et al., 2018), and Recurrent ANNs in SignalP 5.0 (Almagro Armenteros et al., 2019), have demonstrated to be more effective in classifying SP compared to SVM and ANN. Furthermore, the recent SignalP 6.0 based on pLMs, succeeded to classify all five known types of SPs with a higher MCC than its predecessor.

References

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

Bošnjak,I. *et al.* (2014) Occurrence of protein disulfide bonds in different domains of life: a comparison of proteins from the Protein Data Bank. *Protein Eng. Des. Sel.*, **27**, 65–72.

Boser,B.E. et al. (1992) A training algorithm for optimal margin classifiers. Associ. ation for Computing Machinery,pp. 144–152.

Breiman,L. (2001) Random Forests. *Mach. Learn.*, **45**, 5–32.

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

Crooks,G.E. *et al.* (2004) WebLogo: A Sequence Logo Generator. *Genome Res.*, **14**, 1188–1190.

Dyrløv Bendtsen,J. *et al.* (2004) Improved Prediction of Signal Peptides: SignalP 3.0. *J. Mol. Biol.*, **340**, 783–795.

Emanuelsson,O. *et al.* (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.*, **2**, 953–971.

Kazemian,H.B. *et al.* (2014) Signal peptide discrimination and cleavage site identification using SVM and NN. *Comput. Biol. Med.*, **45**, 98–110.

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.

Nugent,T. and Jones,D.T. (2009) Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics*, **10**, 159.

Owji,H. *et al.* (2018) A comprehensive review of signal peptides: Structure, roles, and applications. *Eur. J. Cell Biol.*, **97**, 422–441.

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.

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

von Heijne,G. (1984) How signal sequences maintain cleavage specificity. *J. Mol. Biol.*, **173**, 243–251.

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