International Journal of
*Molecular Sciences*

MDPI

*Article*

# Prediction of Signal Peptides in Proteins from Malaria Parasites

**Michał Burdukiewicz [1], Piotr Sobczyk [2], Jarosław Chilimoniuk [3], Przemysław Gagat [3] and Paweł Mackiewicz [3,*]**

[1] Faculty of Mathematics and Information Science, Warsaw University of Technology, 00-661 Warszawa, Poland; michalburdukiewicz@gmail.com

[2] Department of Mathematics, Wrocław University of Technology, 50-370 Wrocław, Poland; Piotr.Sobczyk@pwr.edu.pl

[3] Department of Genomics, University of Wrocław, 50-383 Wrocław, Poland; jaroslaw.chilimoniuk@gmail.com (J.C.); przemyslaw.gagat@uwr.edu.pl (P.G.)

* Correspondence: pamac@smorfland.uni.wroc.pl

Received: 23 October 2018; Accepted: 17 November 2018; Published: 22 November 2018

**Abstract:** Signal peptides are N-terminal presequences responsible for targeting proteins to the endomembrane system, and subsequent subcellular or extracellular compartments, and consequently condition their proper function. The significance of signal peptides stimulates development of new computational methods for their detection. These methods employ learning systems trained on datasets comprising signal peptides from different types of proteins and taxonomic groups. As a result, the accuracy of predictions are high in the case of signal peptides that are well-represented in databases, but might be low in other, atypical cases. Such atypical signal peptides are present in proteins found in apicomplexan parasites, causative agents of malaria and toxoplasmosis. Apicomplexan proteins have a unique amino acid composition due to their AT-biased genomes. Therefore, we designed a new, more flexible and universal probabilistic model for recognition of atypical eukaryotic signal peptides. Our approach called signalHsmm includes knowledge about the structure of signal peptides and physicochemical properties of amino acids. It is able to recognize signal peptides from the malaria parasites and related species more accurately than popular programs. Moreover, it is still universal enough to provide prediction of other signal peptides on par with the best preforming predictors.

**Keywords:** apicomplexa; plasmodium; malaria; HSMM; hidden semi-Markov model; signal peptides

## 1. Introduction

### 1.1. Roles and Features of Signal Peptides

Eukaryotic proteins encoded in the nuclear genome are synthesized on free ribosomes in the cytosol or on ribosomes attached to the endoplasmic reticulum, and are subsequently transported to specific subcellular or extracellular compartments. The appropriate localization of a protein is essential for its proper function, and this information is contained in the protein as a short amino acid sequence called a targeting or sorting signal.

An N-terminal signal peptide (SP) is responsible for targeting proteins to the endomembrane system, including the endoplasmic reticulum and the Golgi apparatus, in which proteins undergo folding and posttranslational modifications such as glycosylation and phosphorylation [1]. The SP-bearing proteins can either stay inside these compartments, be inserted into cellular membranes or exported outside the cell. Proteins equipped with SPs play a crucial role in metabolism

($\beta$-galactosidase, pepsins) [2], maintenance of tissue structure (collagen) [3], immune response (interferons, interleukins) [4] and regulation of other organismal functions (prolactin, glucagon) [5].

Despite the low sequence similarity [6], a general structure of the SP was proposed and it includes three main parts: n-region, h-region and c-region (Figure 1) [7,8]. The SP starts with the n-region that is composed of a positively charged stretch of 5–8 amino acid residues. This part probably enforces the proper topology on a polypeptide during its translocation through the endoplasmic reticulum membrane based on the positive-inside rule [9]. The n-region is followed by the h-region, a stretch of 8–12 hydrophobic amino acids, which constitutes the core of the SP and usually forms an $\alpha$-helix. The third component of the SP is the polar and uncharged c-region. This part is usually six residues long and ends with a cleavage site, at which a signal peptidase cuts the SP off during or after protein translocation into the lumen of the endoplasmic reticulum [10]. The cleavage site is characterized by a variable amino acid composition but typically contains small and neutral residues at $-3$ and $-1$ positions [11]. The cleavage site is, however, absent from some membrane proteins, in which the SP also acts as the first transmembrane domain. Such an uncleavable SP is sometimes referred as a signal anchor [12,13]. The amino acid composition and the length of these three characteristic regions vary between SPs and they influence the efficiency of protein translocation [14].

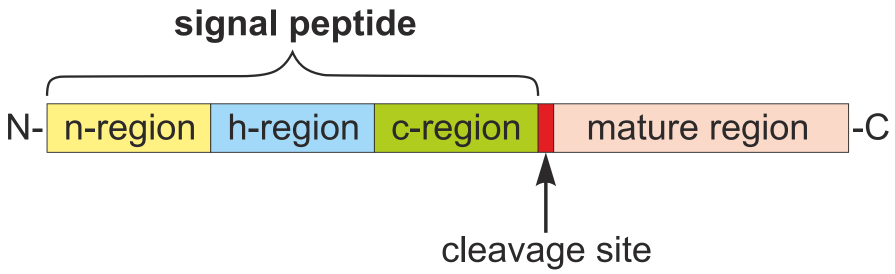

**Figure 1.** Organization of a typical signal peptide (SP). The lengths of SP regions are not drawn in scale.

Understanding the structure of SPs and improving their prediction is of great importance in search of novel drugs and therapies. Equipping proteins with appropriate SPs or their modification can influence protein targeting to various subcellular or extracellular compartments, including substantial increase in their secretion [4,15]. SPs may also participate in tumour immunity, for example the presequence from midkine (a protein contributing to tumour progression) contains epitopes that are recognized by CD4+ T cells [16]. SPs can be potential medication targets in particular for malaria parasites belonging to the genus *Plasmodium* (family Plasmodiidae, phylum Apicomplexa) [17]. Their SPs differ in amino acid composition from typical SPs due to the strongly AT-biased genomes [18]. As a result, we can design therapies that target only these unique presequences and avoid interference with human host SPs. By acting on the malarian SPs, we can disturb many metabolic processes as the plasmodial SPs not only direct proteins into the endomembrane system and outside the cell but also into a digestive vacuole and a unique organelle characteristic only of Apicomplexa, i.e., apicoplast [19–22]. The apicoplast is a reduced four-membrane plastid that lost the photosynthetic function but still plays an important role in synthesis of lipids, heam and iron-sulphur clusters [20,21,23]. These Apicomplexa-specific compartments and their metabolic pathways are further potential targets for antimalarial drugs [24–27], which are intensively searched for due to increasing resistance of malaria parasites to the current drugs [28–32]. However, the existing software for SP prediction, trained and tested on proteins well-represented in available databases, does not perform satisfactorily in the case of Apicomplexa proteins.

### *1.2. Software Predicting Signal Peptides*

Since experimental methods for identification of targeting sequences are time-consuming and laborious, different computational approaches predicting targeting signals were developed.

The software for SP prediction incorporates 'black-box' models, such as: neural networks [33], support vector machines [34], Bayesian networks [35] or k-nearest neighbours [36], for which the decision rules are unknown to the user. More transparent algorithms are based on position matrices or their variants [34,37]. Others, e.g., Phobius, Philius and SignalP 3.0, use hidden Markov models (HMMs) [38–40] that try to reflect the structure of SP regions in their limited probabilistic frameworks. HMMs, however, imply a geometric distribution of the lengths of SP regions. Interestingly, we studied the distribution for the SP regions from the first work applying HMMs in SP prediction [41] and found that the length distribution for each region was not geometric (Figure 2).

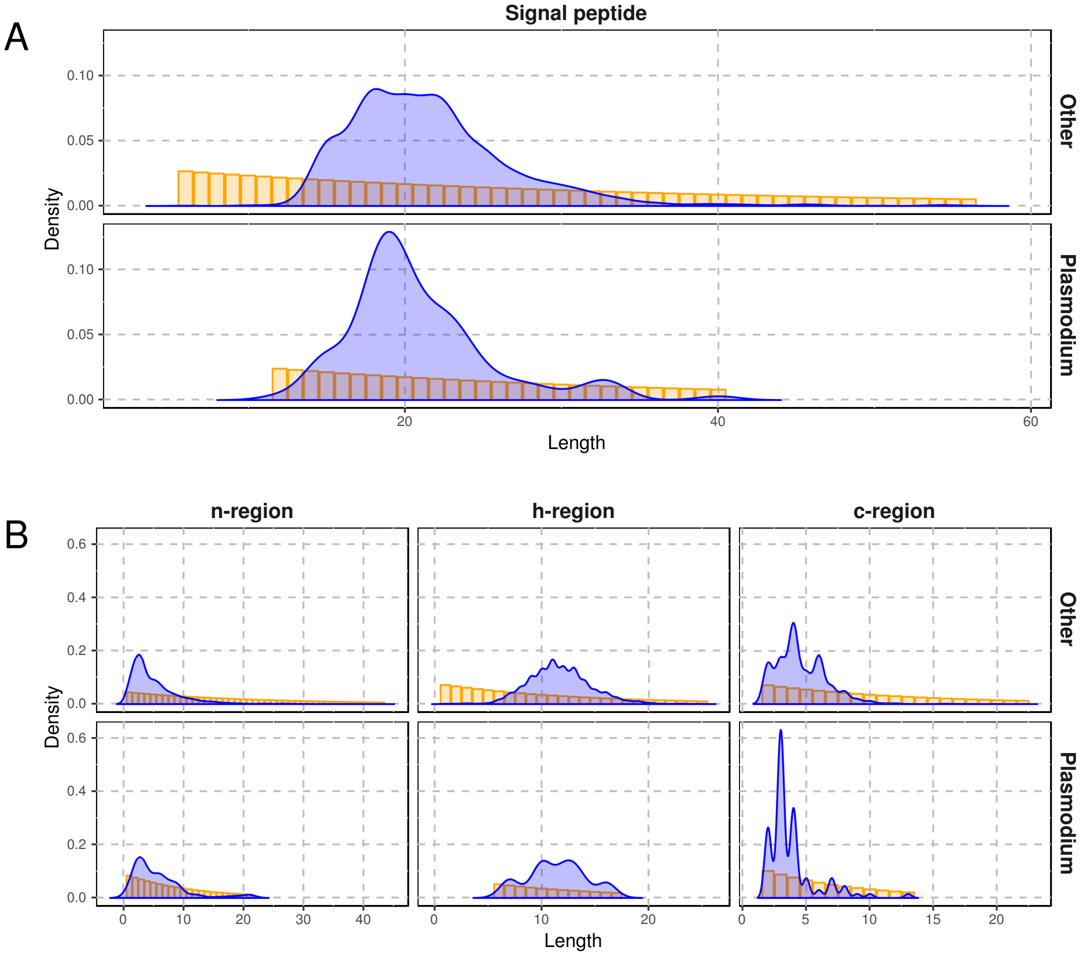

**Figure 2.** Distribution of lengths for SPs (**A**) and their regions (**B**) expressed in the number of amino acid residues for sequences with the representatives of the *Plasmodiidae* family (Plasmodium) and without them (Other). Yellow bars represent a fitted geometric distribution.

The majority of SP predicting software uses the full amino acid alphabet of typical 20 residues. However, this approach treats all residues as separate entities and does not take into account the similarities between amino acids in terms of physicochemical properties. It may result in harmful oversimplification as the SP regions are in fact characterized by specific features of amino acid residues and not by the occurrence of particular amino acids. The only exception is BLOMAP [42], which uses a reduced alphabet of amino acids based on substitution matrices.

Moreover, the existing algorithms require a large number of sequences to be successfully learned. Although the learning sets are constructed to be as diverse as possible, the most frequent sequences dominate the creation of the decision rules. Consequently, the predictors usually accurately identify the most common SPs, but are ineffective in more atypical cases that are less represented in the training datasets. The commonly used rigid scheme of the SP organization (Figure 1) also does not characterize extremely long or short presequences, which constitute a substantial fraction of all SPs.

Therefore, we elaborated a new approach called signalHsmm. In order to enable the prediction of SPs with atypical amino acid composition, as in *Plasmodiidae*, we grouped amino acids based on their physicochemical properties. Instead of looking for patterns consisting of specific residues, we focused on more general features that are necessary for SPs to function properly. This approach was supported by the recent advancements in proteomics suggesting that the reduction of the amino acid alphabet could lead to better fold recognition [43,44]. Considering that one of the key features of SPs is $\alpha$-helix, the use of a shorter amino acid alphabet may indeed improve the SP recognition.

In addition to the alphabet reduction, our algorithm is also based on hidden semi-Markov models. This flexible probabilistic framework does not assume the geometric distribution of the lengths of SP regions, but rather learns the distribution from the training dataset. Therefore, signalHsmm with hidden semi-Markov models and the reduced amino acid alphabet should be more general than its counterparts.

## 2. Results and Discussion

### 2.1. Performance of SignalHsmm Algorithm

In order to evaluate the efficiency of our algorithm, we calculated four performance measures after 5-fold cross-validation procedure for all amino acid encodings: specificity, sensitivity, Matthew's Correlation Coefficient ($\phi$ coefficient) and Area Under the Curve (AUC). The measures were characterized by very small variance (see for example Table 1), which indicates the credibility of the applied cross-validation with 60 repetitions. All the encodings of amino acids showed very good and quite narrow range of AUC (0.93–0.97) and specificity (0.92–0.96). However, the sensitivity was characterized by much wider variation and ranged from 0.66 to 0.94 (Figure 3). For the final signalHsmm algorithm, we selected the encoding that yielded the highest sensitivity and the largest Matthew's Correlation coefficient as well as the second best AUC (Table 2).

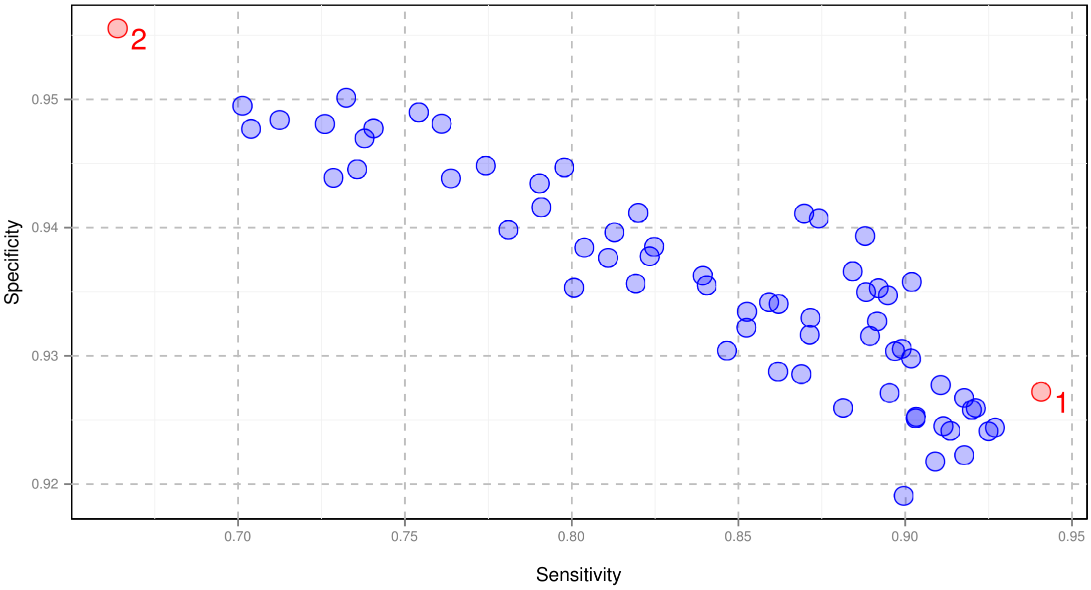

**Figure 3.** Sensitivity and specificity of amino acid encodings after cross-validation. (1) encoding providing the best sensitivity (AUC = 0.9683, MCC = 0.8677), (2) encoding providing the best specificity (AUC = 0.9338, MCC = 0.6474). These encodings are shown in Tables 2 and 3, respectively.

**Table 1.** Performance measures for the amino acid encoding with the highest sensitivity calculated for 60 repetitions of cross-validation.

| Measure | Mean | SD |
|---|---|---|
| AUC | 0.9682 | 0.0023 |
| Sensitivity | 0.9407 | 0.0008 |
| Specificity | 0.9272 | 0.0050 |
| MCC | 0.8681 | 0.0049 |

**Table 2.** The encoding of amino acids with the best sensitivity.

| Group | Amino Acids |
|---|---|
| 1 | D, E, H, K, N, Q, R |
| 2 | G, P, S, T, Y |
| 3 | F, I, L, M, V, W |
| 4 | A, C |

### *2.2. Comparison of Amino Acid Encodings*

We examined in detail the composition of encodings and properties of their amino acids with the best sensitivity and specificity (Tables 2 and 3 and Figure 4). In both cases, group 1 tends to contain generally average-sized polar amino acids. This group is more uniform in the best sensitivity encoding because it includes all charged amino acids, both acidic and basic and also weakly basic histidine, whereas in the best specificity encoding, it does not have histidine, aspartic acid and its amide but contains alanine. These amino acids are nearly absent from the h-region and provide a very good distinction between the SP regions (Figure 4). In the best specificity encoding, for which the polar and charged character of group 1 is not so explicit, the difference in their distribution between the regions is also less visible.

**Table 3.** The encoding of amino acids with the best specificity.

| Group | Amino Acids |
|---|---|
| 1 | A, E, K, Q, R |
| 2 | D, G, N, P, S, T |
| 3 | C, H, I, L, M, V |
| 4 | F, W, Y |

Amino acids belonging to group 2 generally show quite a low probability of occurrence in $\alpha$-helix. The best sensitivity encoding comprises two types of amino acids: all three hydroxylated residues as well as aliphatic glycine and proline known to break $\alpha$-helices. The best specificity encoding lacks tyrosine but additionally includes aspartic acid and its amide, which increase the polar character of the group. Despite these differences, the occurrence of group 2 in both encodings is very similar in all SP regions (Figure 4). This group is the rarest in h-region and the most frequent in c-region.

Both encodings have strongly non-polar and aliphatic amino acids in group 3, such as: isoleucine, leucine, methionine and valine. The hydrophobic property of this group is pronounced in the best sensitivity encoding by the presence of aromatic tryptophan and phenylalanine, whereas the group 3 in the best specificity encoding includes also hydrophobic cysteine and slightly basic but aromatic histidine. Because of the hydrophobic character, this group dominates in the h-region in both amino acid classifications (Figure 4).

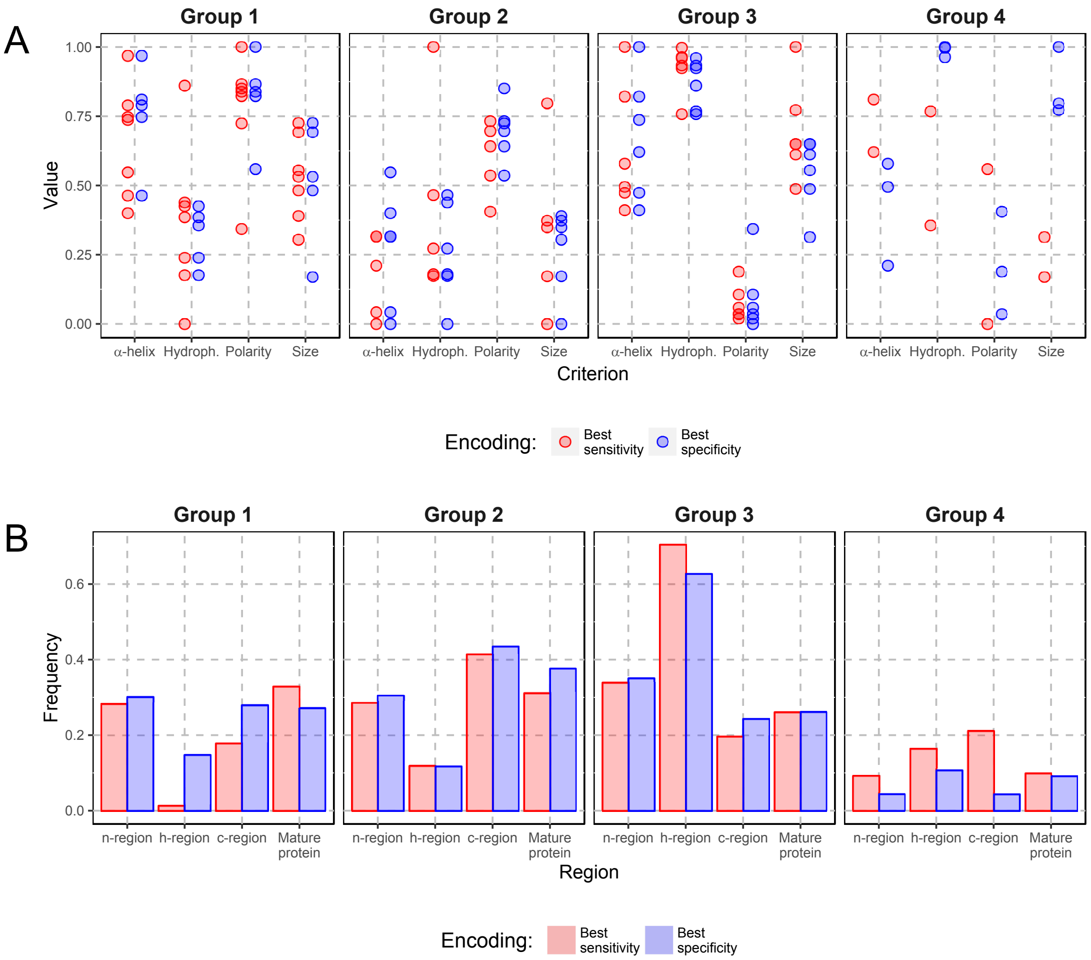

**Figure 4.** Comparison of amino acids classified into four groups providing the best sensitivity and specificity in the SP recognition. Normalized value of properties for particular amino acids represented by points (**A**). Frequencies of amino acids from the four groups in different regions of signal peptide and mature proteins (**B**).

Group 4 is the most diverse in both encodings. In the case of the best sensitivity encoding, this group comprises only alanine and cysteine, which are rather small amino acids and tend to appear in $\alpha$-helices. This very unique composition seems to be the most typical of the SP c-region. In contrast, group 4 in the best specificity encoding contains large aromatic amino acids: phenylalanine, tryptophan and tyrosine without special preference to the SP regions.

The encodings of amino acids play a crucial role in the recognition of SPs but do not affect to such an extent the identification of proteins without the SP. The change in the specificity for different encodings is seven times smaller than for sensitivity (Figure 3). It results from more uniform distribution of different residues in the mature protein than in the SP regions.

### *2.3. Benchmark Tests*

In order to provide fair comparison of our algorithm with the previous software, we trained our model on 2311 SP-bearing sequences deposited in Uniprot until 2010 (the iteration of signalHsmm called signalHsmm-2010) (Figure 5). The set corresponded to the data used to train SignalP 4.1, the newest classifier present in the benchmark. In addition, we prepared a smaller dataset, covering only 336 sequences collected until 1987, which is the year the first method predicting SPs was published [45]. The signalHsmm-1987 iteration had to extract the SP model from the dataset more limited than the training sets used by all the classifiers included in the benchmark.

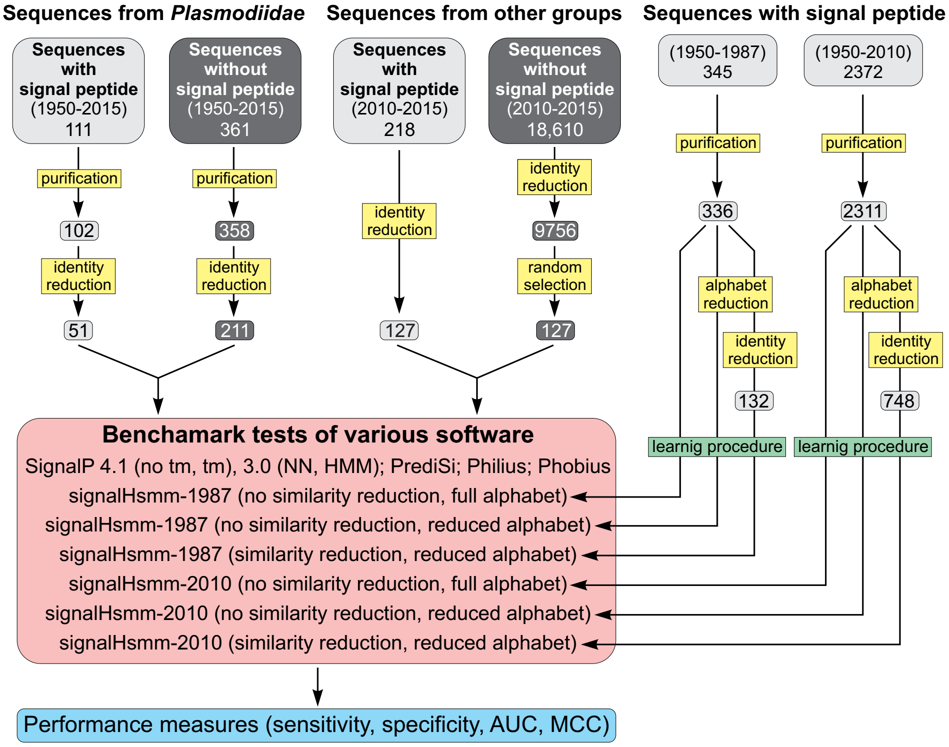
Sequences from *Plasmodiidae*
Sequences from other groups
Sequences with signal peptide
Sequences with signal peptide (1950-2015) 111
Sequences without signal peptide (1950-2015) 361
Sequences with signal peptide (2010-2015) 218
Sequences without signal peptide (2010-2015) 18,610
(1950-1987) 345
(1950-2010) 2372
purification
102
identity reduction
51
purification
358
identity reduction
211
identity reduction
127
identity reduction
9756
random selection
127
purification
336
alphabet reduction
identity reduction
132
purification
2311
alphabet reduction
identity reduction
748
learnig procedure
learnig procedure
Benchamark tests of various software
SignalP 4.1 (no tm, tm), 3.0 (NN, HMM); PrediSi; Philius; Phobius
signalHsmm-1987 (no similarity reduction, full alphabet)
signalHsmm-1987 (no similarity reduction, reduced alphabet)
signalHsmm-1987 (similarity reduction, reduced alphabet)
signalHsmm-2010 (no similarity reduction, full alphabet)
signalHsmm-2010 (no similarity reduction, reduced alphabet)
signalHsmm-2010 (similarity reduction, reduced alphabet)
Performance measures (sensitivity, specificity, AUC, MCC)

**Figure 5.** Data selection, training, testing and evaluation of signalHsmm.

Together with signalHsmm, we evaluated several SP predicting algorithms in terms of recognition of atypical SPs from malaria parasites: SignalP 4.1, PrediSi, Phobius and Philius (see Table 4 for the most common performance measures, and Supplemental Table S1 for 23 performance measures). The older version of SignalP, i.e., SignalP 3.0, was also incorporated in the analysis because it is often chosen over SignalP 4.1 in the analysis of sequences belonging to the Apicomplexa due to its larger sensitivity [46,47]. In this comparison, our algorithm obtained the greatest AUC, MCC and maximal sensitivity. For 15 of 23 performance measures, signalHsmm was the best of all (Supplemental Table S1). PrediSi received the best specificity but at the expense of significantly reduced sensitivity.

**Table 4.** Comparison of Sensitivity, Specificity, Matthews Correlation Coefficient (MCC) and Area Under the Curve (AUC) for different classifiers and signal peptide-bearing proteins from members of *Plasmodiidae*. The abbreviations ’tm’ and ’no tm’ indicate version considering and not considering transmembrane domains, whereas ’ident. 50%’ means removal from the learning set sequences with the sequence identity larger than 50%.

| | **Sensitivity** | **Specificity** | **MCC** | **AUC** |
|---|---|---|---|---|
| SignalP 4.1 (no tm) [33] | 0.8235 | 0.9100 | 0.6872 | 0.8667 |
| SignalP 4.1 (tm) [33] | 0.6471 | 0.9431 | 0.6196 | 0.7951 |
| SignalP 3.0 (NN) [40] | 0.8824 | 0.9052 | 0.7220 | 0.8938 |
| SignalP 3.0 (HMM) [40] | 0.6275 | 0.9194 | 0.5553 | 0.7734 |
| PrediSi [37] | 0.3333 | **0.9573** | 0.3849 | 0.6453 |
| Philius [39] | 0.6078 | 0.9336 | 0.5684 | 0.7707 |
| Phobius [38] | 0.6471 | 0.9289 | 0.5895 | 0.7880 |
| signalHsmm-2010 | 0.9804 | 0.8720 | 0.7409 | 0.9262 |
| signalHsmm-2010 (ident. 50%) | **1.0000** | 0.8768 | **0.7621** | **0.9384** |
| signalHsmm-2010 (raw aa) | 0.8431 | 0.9005 | 0.6853 | 0.8718 |
| signalHsmm-1987 | 0.9216 | 0.8910 | 0.7271 | 0.9063 |
| signalHsmm-1987 (ident. 50%) | 0.9412 | 0.8768 | 0.7194 | 0.9090 |
| signalHsmm-1987 (raw aa) | 0.7647 | 0.9052 | 0.6350 | 0.8350 |

We trained several iterations of signalHsmm described above to check improvements introduced to our software, i.e., the new probabilistic model (hidden semi-Markov model) and the simplified alphabet of amino acids. The latter was compared with the version trained on raw amino acid sequences, denoted as 'raw aa' in Table 4. The performance of this iteration was mostly worse than the performance of its counterpart trained on the reduced alphabet. The version with the amino acid encodings outperformed the version with simple amino acids in 17 of 23 measures (Supplemental Table S1). The obtained results confirm that the function of the SP does not depend on specific amino acids but on more general features, and our simpler model is able to recognize the unique structure of the SP more accurately. The advantage of the hidden semi-Markov model over normal Markov model was presented through the comparison of signalHsmm with SP predictors utilizing HMM: Phobius, Philius and SignalP 3.0 (HMM)—Table 4. For 17 of 23 measures including AUC, MCC and sensitivity, our algorithm outperformed all of them (Table 4, Supplemental Table S1).

In order to check the susceptibility of our model to overfitting, we trained it on a dataset with 50%-sequence identity reduction. We discovered that our model, probably thanks to its relative simplicity, did not overfit and versions trained on the set with and without the restrictive sequence identity reduction were comparable. Moreover, the former was slightly better in 21 measures than the model based on all sequences.

The overall simplicity of our approach does not hinder its capabilities to recognise SPs from other organisms. We benchmarked signalHsmm iterations and other software on the set of 127 eukaryotic proteins with and without SPs that were added to the UniProt database after 2010. Their sequence identity in the testing set was also reduced as described above. SignalHsmm-2010 performed comparably to SignalP (see Supplemental Table S2 for all performance measures). Its AUC was 0.94 compared to 0.95 and 0.96 of the two SignalP versions. For 20 measures, it was the second in the ranking just after SignalP programs and, for five parameters (sensitivity, recall, true positive rate, the number of true positives and false negatives), it outperformed the newest version of SignalP.

### 2.4. Specific Composition of Plasmodiidae Signal Peptides

The plasmodial genomes are characterized by a large excess of adenine and thymine [18], and this strongly influences the amino acid composition of their encoded proteins including SPs (Figure 6). SPs differ significantly between *Plasmodiidae* and other taxa in composition of 13 amino acids (Wilcoxon test, $p < 0.05$ corrected by the Benjamini–Hochberg method in the multiple testing). The plasmodial SPs are especially abundant in amino acids coded by codons rich in adenine and thymine, such as: phenylalanine (TTY), isoleucine (ATH), lysine (AAR) and asparagine (AAY), whereas they are poor in amino acids coded by guanine and cytosine rich codons: alanine (GCN), glycine (GGN) and proline (CCN). Leucine is coded by a set of codons with mixed composition (TTR, CTN) but also discriminates the two sets of SPs.

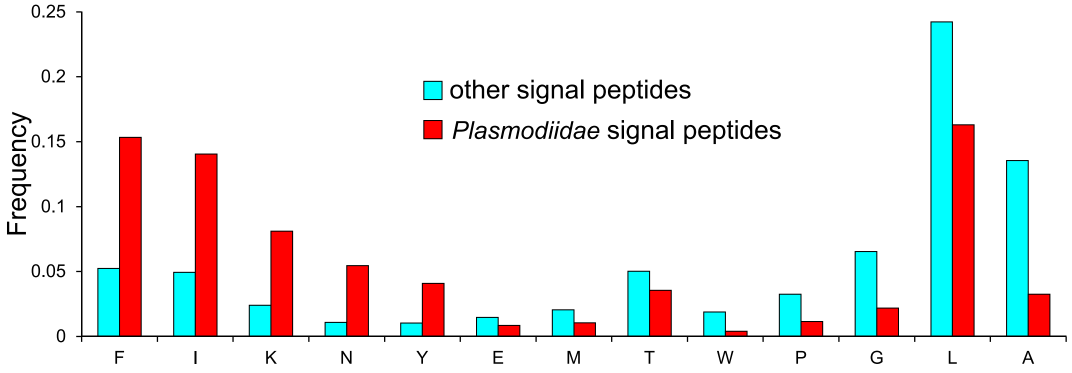

**Figure 6.** Mean frequency of amino acids that significantly discriminate SPs from *Plasmodiidae* and other organisms. Amino acids were arranged according to the difference in the mean values.

As a result, plasmodial SPs separate from other SPs according to raw amino acid composition in Principal Component Analysis (Figure 7A). Mature proteins of *Plasmodiidae* are also shifted in the plot from other mature proteins. Therefore, algorithms that consider only particular amino acids may not have decision rules appropriate for such composition and consequently fail to identify SPs. Interestingly, the amino acid encoding employed by signalHsmm reduces this difference and unifies all SPs into one set (Figure 7B). Similarly, when the reduced amino acid alphabet is considered, mature proteins are also inseparable in the plot. It should be emphasized that the reduction of the amino acid alphabet used by our algorithm does not weaken the difference between SPs and mature proteins, which still create distinguishable groups. These analyses indicate that the applied amino acid encoding enables capturing a general composition of many SPs keeping simultaneously their difference from mature proteins. The use of physicochemical properties of amino acids instead of raw sequences also improved the prediction of protein function in four representative model organisms: *Homo sapiens*, *Arabidopsis thaliana*, *Saccharomyces cerevisia*e and *Mycobacterium tuberculosis* [48].

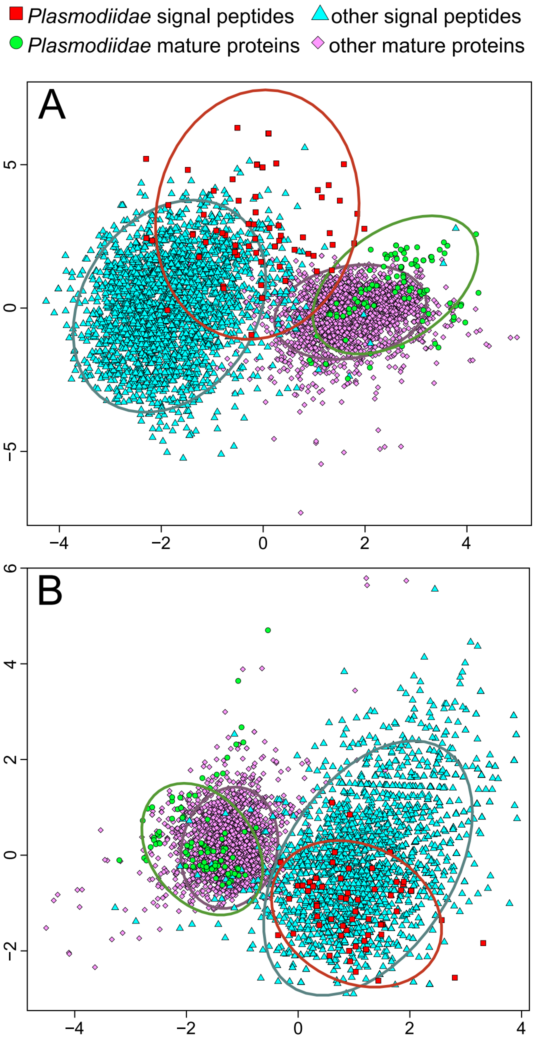

**Figure 7.** Principal Component Analysis performed on raw amino acid composition (**A**) and amino acids encoding into groups (**B**) for SPs and mature proteins from *Plasmodiidae* and other organisms. SPs from *Plasmodiidae* create a set, which is quite distinct from other SPs according to raw amino acid composition. However, the application of amino acid encoding chosen in cross-validation makes these sets similar but still supporting a significant difference between SPs and mature proteins. Concentration ellipses cover 95% of points from a given set.

## 3. Materials and Methods

### *3.1. Overview*

Since the functionality of SPs depends on the physicochemical properties of residues in a given SP region, we clustered amino acids into several groups based on this criterion. The pre-processed sequences were further analysed by an enhanced version of a heuristic algorithm employed in SignalP 2.0, which determines borders between the three SP regions [41]. We also refined some criteria for recognition of the regions to attune the algorithm to atypical SPs. Next, two models were trained to detect proteins with and without the SP. The first one was a hidden semi-Markov model, in which each of the three SP regions was represented by a different hidden state. The additional fourth hidden state represented the mature protein. Each state was described by its frequencies of amino acid groups. The distribution of hidden states durations, i.e., the number of amino acids, was based on the empirical distribution of region lengths from the training set. Furthermore, the hidden semi-Markov model was enriched with n-grams representing SP cleavage sites. The second model was a simple probabilistic approach in which no association between amino acids was assumed and the probability of presence of amino acids groups was determined by their frequencies in mature proteins.

### *3.2. Data Selection*

The final predictor was trained on 2438 experimentally confirmed SPs from eukaryotic proteins, which sequences and annotations were downloaded from the UniProt database release 2015_06. We removed sequences with more than one cleavage site, an unknown cleavage site and ambiguous symbols of amino acid residues: X, J, Z, B and U (selenocysteine). Sequences without the SP were randomly selected in the same number as the positive set. Moreover, we created a learning subset of 2311 sequences deposited in the database until 2010 to perform fair comparison of our algorithm with older software that was trained on smaller number of sequences (Figure 5). We also used a subset of 336 sequences present in the database until 1987, which is the year the first method predicting the SP was published, to check susceptibility of our algorithm to a limited amount of information.

The main testing set consisted of proteins from the *Plasmodiidae* family (Figure 5). The positive set contained 102 sequences with a putative SP with the annotated start and cleavage site. The corresponding negative set comprised 358 sequences without any SP information. The other testing set consisted of 127 eukaryotic proteins with an SP included in the UniProt database after 2010.

### *3.3. Sequence Identity Reduction in Studied Sequence Sets*

In order to reduce the set according to identity of collected protein sequences, we filtered them using cd-hit [49]. SP sequences and the first 70 amino acid residues of proteins without the peptide were subjected to sequence identity reduction according to Nielsen et al. [50]. We prepared two reduced learning datasets with sequences deposited until 2010 and 1987 by removing sequences with identity larger than 50% threshold (word length 2). After this procedure, the sets contained 748 and 132 sequences with SPs, respectively. The main *Plasmodiidae* set was filtered in the same manner and reduced to 51 and 211 sequences with and without the SP, respectively.

### *3.4. Clustering of Amino Acids into Groups*

In order to reduce the alphabet of amino acids, we clustered them into several physicochemical groups, essentially re-using our methodology from a previous study [51]. It is a different approach compared to BLOMAP [42], which also uses a reduced alphabet of amino acids, but based on substitution matrices. We grouped amino acids using four properties relevant for the structure of the SP: their hydrophobicity, tendency to build $\alpha$-helices, polarity and size. The high hydrophobicity is a good determinant of the h-region, which $\alpha$-helix secondary structure is probably induced by the positively charged n-region. The high polarity as well as small size are important features of residues in the c-region and cleavage site [11].

**Table 5.** Properties of amino acids used in their clusterization.

| Property Name | Amino Acid Scale |
|---|---|
| Size | Size [52] |
| Size | Molecular weight [53] |
| Size | Residue volume [54] |
| Size | Bulkiness [55] |
| Hydrophobicity | Normalized hydrophobicity scales for $\alpha$-proteins [56] |
| Hydrophobicity | Consensus normalized hydrophobicity scale [57] |
| Hydrophobicity | Hydropathy index [58] |
| Hydrophobicity | Surrounding hydrophobicity in $\alpha$-helix [59] |
| Polarity | Polarity [60] |
| Polarity | Mean polarity [61] |
| Occurrence in $\alpha$-helices | Signal sequence helical potential [62] |
| Occurrence in $\alpha$-helices | Normalized frequency of N-terminal helix [63] |
| Occurrence in $\alpha$-helices | Relative frequency in $\alpha$-helix [64] |

We considered in total 13 amino acid scales present in AAIndex database [65] (Table 5). We selected one scale per a given property and carried out all possible 96 permutations of them. Based on that, we created 96 possible clusterings of amino acids using Euclidean distance and Ward's method. Next, we cut the clusterings to create four groups of amino acids. In 31% of cases, the groupings were identical. To compare the usefulness of these encodings, we performed a 5-fold cross-validation training of our algorithm on every encoding. We created balanced data sets by subsampling proteins without the SP to equal the number of proteins with the SP. The cross-validation was repeated 60 times to ensure that every protein without the SP was included in the learning set with the probability higher than 0.5.

### 3.5. Hidden Semi-Markov Model

Our algorithm is based on a hidden semi-Markov model (HSMM), which is an extension of the hidden Markov model (HMM) [66–68]. The HMM consists of two stochastic processes. The first is a discrete Markov chain $X_{t=1}^{T}$ on the set of hidden states $\{S_1, \dots S_n\}$, where $t$ means a step of this process and $T$ means total duration of the process corresponding to the length of the SP. Hidden states represent particular SP regions and are 'the cause' of the observations, which are amino acid residues in analysed sequences. In a subsequent step $t+1$, the hidden state might change to another according to a transition matrix $A = (a)_{i,j=1}^{n}$, where $a_{i,j} = \mathcal{P}(X_{t+1} = S_j | X_t = S_i)$ means a probability of being in a state $j$ on condition that in the previous step was a state $i$. The second process $E_{t=1}^{T}$ is an observation process defined on the set of possible observations $\{O_1, \dots, O_m\}$. They are assumed to occur independently but conditionally on the hidden states that emits these observations. Distribution of the observations are given by a matrix $B = (b)_{i,k=1}^{m}$, where $b_{i,k} = \mathcal{P}(E_t = O_k | X_t = S_i)$ means a probability of emission of observation $k$ on condition that a hidden state was $i$. The main goal of signalHsmm is to find the most probable SP regions boundaries for a given sequence. This is achieved with the Viterbi algorithm.

In the regular HMM, the hidden state duration, i.e., the number of observations emitted by the hidden state, has a geometric distribution. Durbin et al. [69] showed how to extend it for different distributions without significant increase in computational complexity. Similar ideas were used for SP recognition, for example by Käll et al. [38]. However, it is still not flexible enough because the empirical regional length distributions (see Figure 2) are difficult to capture in this way.

In our approach, we used a modification of HMM called hidden semi-Markov model (HSMM) [67]. It extends the HMM by assuming a duration distribution for a given hidden state (Figure 8). Then, the model includes additionally probabilities of duration in hidden states:

$$\mathcal{P}(\text{duration in state} = d | \text{state is } S_i), \quad i = 1, \dots, n, \quad d = 1, \dots, D,$$

where $D$ is the maximum allowed duration. Since our data sets are sufficiently large and $D$ is small—around 30 amino acid residues, computational effort is not much higher than in the regular HMM.

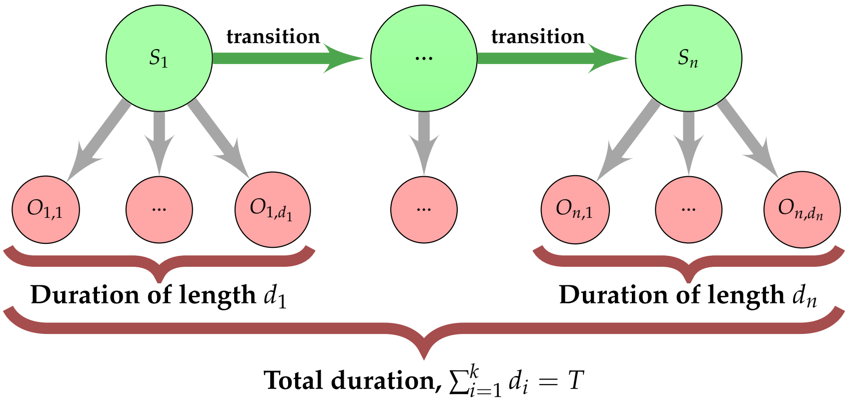

**Figure 8.** General scheme of hidden semi-Markov model. The model consists of hidden states $S$ representing regions of the SP with the length distribution given by the state duration $d$. The hidden states emit observations $O$ which are amino acid residues.

Almost all entries in the transition matrix $A$ are zeros because regions represented by hidden states are sequential. The possible transitions between them are depicted as arrows in Figure 9. The probabilities of observations for the hidden states and hidden states durations were estimated from the training data. The advantage of the HSMM model results not only from its better performance but also form its straightforwardness and flexibility.

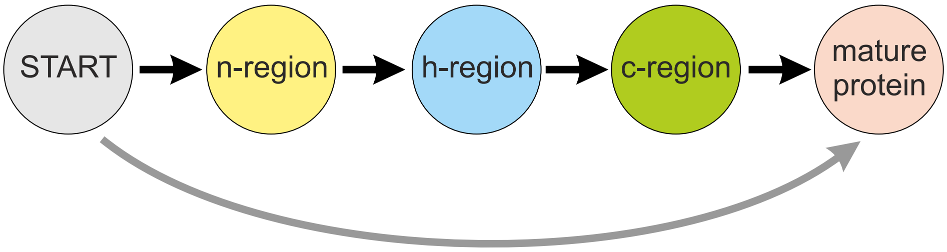

**Figure 9.** Diagram of signalHsmm showing hidden states which represent the SP regions and mature protein. Arrows indicate transitions between these states.

## 4. Conclusions

We proposed a novel solution to the problem of SP prediction, which is very efficient in recognition of atypical SPs from plasmodial proteins despite the fact that the program was trained on data coming from all eukaryotes. It indicates that our algorithm is able to describe common features of all SPs based on the classical division of the SP into three regions. Our software is not limited to very specific taxonomic group, and is able to compete with state-of-art algorithms in detecting SPs of other organisms.

One of the most important features of signalHsmm is its stability. The difference in performance for versions trained on large and small datasets deposited in databases at different times is negligible. It implies that signalHsmm, thanks to its unique structure, extracts roughly the same general information from SPs regardless of the size and type of the training dataset. Similarly, iterations trained on datasets with and without the removal of redundancy, resulting from sequences homology, showed similar prediction efficiency. For the first dataset, the algorithm was even slightly better. It suggests that our probabilistic model is quite resistant to overfitting and does not adjust itself to the most common patterns in the training dataset but retrieves the universal SP model.

The existing software detecting SPs does not usually reveal decision rules responsible for the prediction. Our algorithm is the first step to explicitly show features of SPs important in their recognition, which is interesting from the biological point of view. The applied encoding of amino acids not only reduces the dimensionality of the problem, but also makes our probabilistic model more interpretable. Thanks to that, we were able to determine physicochemical properties of amino acids for particular SP regions. Our model confirmed not only the high hydrophobicity of the h-region and polarity of the n-region but also found that hydroxylated amino acids are one of the most typical amino acids in the c-region. In contrast to the h-region, it also contains $\alpha$-helix breakers: glycine and proline.

The flexibility and efficiency in recovering information makes signalHsmm unique among similar software. It properly models very specific SPs belonging to narrow taxonomic groups that are poorly represented in databases and can effectively extract information from very small datasets. Our approach may lead in future to development of new predictors specialized in recognition of atypical signals targeting sequences to variuos subcellular compartments.

The prediction of SPs is of great importantance as proteins equipped with these targeting signals are involved in many processes associated with diseases and disorders. Therefore, better knowledge and prediction of SPs can enable scientists to discover new drugs and development of more efficient therapies. Particularly, good drug targets could be proteins containing SPs that are targeted to plasmodial specific compartments such as apicoplasts.

**Supplementary Materials:** Supplementary materials can be found at http://www.mdpi.com/1422-0067/19/12/3709/s1. Table S1: **Benchmark results—*Plasmodiidae***. Performance measures for benchmark test of SP predictors using sequences belonging to *Plasmodiidae*. Table S2: **Benchmark results—Eukaryotes**. Performance measures for benchmark test of SP predictors using all eukaryotic sequences. The signalHsmm prediction web-server is available at: http://smorfland.uni.wroc.pl/shiny/signalHsmm/. SignalHsmm is implemented as an R package available at: https://cran.r-project.org/package=signalHsmm. The stand-alone version offers prediction and tools to build, train and test novel SP models.

**Author Contributions:** All authors participated in the design of the study. M.B. and P.S. created the probabilistic model. M.B. and P.S. implemented the model and wrote the software. M.B., P.M. and J.C. analysed the data. M.B., P.G. and P.M. wrote the manuscript. All authors critically revised the manuscript.

**Funding:** This research was funded by the National Science Centre grant UMO-2017/26/D/NZ8/00444 to P.G. and M.B. as well as grants 2015/17/N/NZ2/01845 and 2017/24/T/NZ2/00003 to M.B. Publication costs were supported by Wroclaw Center of Biotechnology program "The Leading National Research Center (KNOW) for years 2014–2018".

**Acknowledgments:** We want to thank Paweł Błażej (University of Wrocław) for fruitful discussion about hidden semi-Markov models.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| SP | Signal Peptide |
| HSMM | Hidden Semi-Markov Model |
| HMM | Hidden Markov Model |
| AUC | Area Under the Curve |
| MCC | Matthews Correlation Coefficient |

## References

1. Rapoport, T.A. Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. *Nature* **2007**, *450*, 663–669. [CrossRef] [PubMed]
2. Hofmann, K.J.; Schultz, L.D. Mutations of the alpha-galactosidase signal peptide which greatly enhance secretion of heterologous proteins by yeast. *Gene* **1991**, *101*, 105–111. [CrossRef]

3. Chan, D.; Ho, M.S.P.; Cheah, K.S.E. Aberrant signal peptide cleavage of collagen X in Schmid metaphyseal chondrodysplasia. Implications for the molecular basis of the disease. *J. Biol. Chem.* **2001**, *276*, 7992–7997. [CrossRef] [PubMed]
4. Zhang, L.; Leng, Q.; Mixson, A.J. Alteration in the IL-2 signal peptide affects secretion of proteins in vitro and in vivo. *J. Gene Med.* **2005**, *7*, 354–365. [CrossRef] [PubMed]
5. Huang, Y.; Wilkinson, G.F.; Willars, G.B. Role of the signal peptide in the synthesis and processing of the glucagon-like peptide-1 receptor. *Br. J. Pharmacol.* **2010**, *159*, 237–251. [CrossRef] [PubMed]
6. Ladunga, I. PHYSEAN: PHYsical SEquence ANalysis for the identification of protein domains on the basis of physical and chemical properties of amino acids. *Bioinformatics* **1999**, *15*, 1028–1038. [CrossRef] [PubMed]
7. Izard, J.W.; Kendall, D.A. Signal peptides: Exquisitely designed transport promoters. *Mol. Microbiol.* **1994**, *13*, 765–773. [CrossRef] [PubMed]
8. Voss, M.; Schröder, B.; Fluhrer, R. Mechanism, specificity, and physiology of signal peptide peptidase (SPP) and SPP-like proteases. *Biochim. Biophys. Acta* **2013**, *1828*, 2828–2839. [CrossRef] [PubMed]
9. von Heijne, G.; Gavel, Y. Topogenic signals in integral membrane proteins. *Eur. J. Biochem.* **1988**, *174*, 671–678. [CrossRef] [PubMed]
10. Paetzel, M.; Karla, A.; Strynadka, N.C.; Dalbey, R.E. Signal peptidases. *Chem. Rev.* **2002**, *102*, 4549–4580. [CrossRef] [PubMed]
11. Palzkill, T.; Le, Q.Q.; Wong, A.; Botstein, D. Selection of functional signal peptide cleavage sites from a library of random sequences. *J. Bacteriol.* **1994**, *176*, 563–568. [CrossRef] [PubMed]
12. Szczesna-Skorupa, E.; Browne, N.; Mead, D.; Kemper, B. Positive charges at the NH2 terminus convert the membrane-anchor signal peptide of cytochrome P-450 to a secretory signal peptide. *Proc. Natl. Acad. Sci. USA* **1988**, *85*, 738–742. [CrossRef] [PubMed]
13. Zhang, P.; Tan, S.; Berry, J.O.; Li, P.; Ren, N.; Li, S.; Yang, G.; Wang, W.B.; Qi, X.T.; Yin, L.P. An Uncleaved signal peptide directs the *Malus xiaojinensis* iron transporter protein Mx IRT1 into the ER for the PM Secretory Pathway. *Int. J. Mol. Sci.* **2014**, *15*, 20413–20433. [CrossRef] [PubMed]
14. Hegde, R.S.; Bernstein, H.D. The surprising complexity of signal sequences. *Trends Biochem. Sci.* **2006**, *31*, 563–571. [CrossRef] [PubMed]
15. Moeller, L.; Taylor-Vokes, R.; Fox, S.; Gan, Q.; Johnson, L.; Wang, K. Wet-milling transgenic maize seed for fraction enrichment of recombinant subunit vaccine. *Biotechnol. Prog.* **2010**, *26*, 458–465. [CrossRef] [PubMed]
16. Kerzerho, J.; Schneider, A.; Favry, E.; Castelli, F.A.; Maillère, B. The signal peptide of the tumor-shared antigen midkine hosts CD4$^+$ T cell epitopes. *J. Biol. Chem.* **2013**, *288*, 13370–13377. [CrossRef] [PubMed]
17. Neto Ade, M.; Alvarenga, D.A.; Rezende, A.M.; Resende, S.S.; Ribeiro Rde, S.; Fontes, C.J.; Carvalho, L.H.; de Brito, C.F. Improving N-terminal protein annotation of *Plasmodium* species based on signal peptide prediction of orthologous proteins. *Malar. J.* **2012**, *11*, 375. [CrossRef] [PubMed]
18. Tonkin, C.J.; Kalanon, M.; McFadden, G.I. Protein targeting to the malaria parasite plastid. *Traffic* **2008**, *9*, 166–175. [CrossRef] [PubMed]
19. Foth, B.J.; McFadden, G.I. The apicoplast: A plastid in *Plasmodium falciparum* and other Apicomplexan parasites. *Int. Rev. Cytol.* **2003**, *224*, 57–110. [PubMed]
20. Lim, L.; McFadden, G.I. The evolution, metabolism and functions of the apicoplast. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **2010**, *365*, 749–763. [CrossRef] [PubMed]
21. McFadden, G.I. The apicoplast. *Protoplasma* **2011**, *248*, 641–650. [CrossRef] [PubMed]
22. Heiny, S.R.; Pautz, S.; Recker, M.; Przyborski, J.M. Protein traffic to the *Plasmodium falciparum* apicoplast: Evidence for a sorting branch point at the Golgi. *Traffic* **2014**, *15*, 1290–1304. [CrossRef] [PubMed]
23. Mazumdar, J.; H Wilson, E.; Masek, K.; A Hunter, C.; Striepen, B. Apicoplast fatty acid synthesis is essential for organelle biogenesis and parasite survival in *Toxoplasma gondii*. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 13192–13197. [CrossRef] [PubMed]
24. Fichera, M.E.; Roos, D.S. A plastid organelle as a drug target in apicomplexan parasites. *Nature* **1997**, *390*, 407–409. [CrossRef] [PubMed]
25. Ralph, S.A.; D'Ombrain, M.C.; McFadden, G.I. The apicoplast as an antimalarial drug target. *Drug. Resist. Updat.* **2001**, *4*, 145–151. [CrossRef] [PubMed]
26. Gornicki, P. Apicoplast fatty acid biosynthesis as a target for medical intervention in apicomplexan parasites. *Int. J. Parasitol.* **2003**, *33*, 885–896. [CrossRef]

27. Garcia-Estrada, C.; Prada, C.F.; Fernandez-Rubio, C.; Rojo-Vazquez, F.; Balana-Fouce, R. DNA topoisomerases in apicomplexan parasites: Promising targets for drug discovery. *Proc. Biol. Sci.* **2010**, *277*, 1777–1787. [CrossRef] [PubMed]
28. Vandomme, A.; Fréville, A.; Cailliau, K.; Kalamou, H.; Bodart, J.F.; Khalife, J.; Pierrot, C. PhosphoTyrosyl phosphatase activator of *Plasmodium falciparum*: Identification of its residues involved in binding to and activation of PP2A. *Int. J. Mol. Sci.* **2014**, *15*, 2431–2453. [CrossRef] [PubMed]
29. Ng, C.L.; Fidock, D.A.; Bogyo, M. Protein degradation systems as antimalarial therapeutic targets. *Trends Parasitol.* **2017**, *33*, 731–743.[CrossRef] [PubMed]
30. Jiménez-Díaz, M.B.; Ebert, D.; Salinas, Y.; Pradhan, A.; Lehane, A.M.; Myrand-Lapierre, M.E.; O'Loughlin, K.G.; Shackleford, D.M.; de Almeida, M.J.; Carrillo, A.K.; et al. (+)-SJ733, a clinical candidate for malaria that acts through ATP4 to induce rapid host-mediated clearance of *Plasmodium*. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, E5455–E5462. [CrossRef] [PubMed]
31. Vaidya, A.B.; Morrisey, J.M.; Zhang, Z.; Das, S.; Daly, T.M.; Otto, T.D.; Spillman, N.J.; Wyvratt, M.; Siegl, P.; Marfurt, J.; et al. Pyrazoleamide compounds are potent antimalarials that target Na+ homeostasis in intraerythrocytic *Plasmodium falciparum*. *Nat. Commun.* **2014**, *5*, 5521. [CrossRef] [PubMed]
32. Phillips, M.A.; Lotharius, J.; Marsh, K.; White, J.; Dayan, A.; White, K.L.; Njoroge, J.W.; El Mazouni, F.; Lao, Y.; Kokkonda, S.; et al. A Long-Duration Dihydroorotate Dehydrogenase Inhibitor (DSM265) for Prevention and Treatment of Malaria. *Sci. Transl. Med.* **2015**, *7*, 296ra111. [CrossRef] [PubMed]
33. Petersen, T.N.; Brunak, S.; von Heijne, G.; Nielsen, H. SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nat. Methods* **2011**, *8*, 785–786. [CrossRef] [PubMed]
34. Zhang, S.W.; Zhang, T.H.; Zhang, J.N.; Huang, Y. Prediction of signal peptide cleavage sites with subsite-coupled and template matching fusion algorithm. *Mol. Inform.* **2014**, *33*, 230–239. [CrossRef] [PubMed]
35. Zheng, Z.; Chen, Y.; Chen, L.; Guo, G.; Fan, Y.; Kong, X. Signal-BNF: A Bayesian network fusing approach to predict signal peptides. *J. Biomed. Biotechnol.* **2012**, *2012*, 492174. [CrossRef] [PubMed]
36. Shen, H.B.; Chou, K.C. Signal-3L: A 3-layer approach for predicting signal peptides. *Biochem. Biophys. Res. Commun.* **2007**, *363*, 297–303. [CrossRef] [PubMed]
37. Hiller, K.; Grote, A.; Scheer, M.; Münch, R.; Jahn, D. PrediSi: Prediction of signal peptides and their cleavage positions. *Nucleic Acids Res.* **2004**, *32*, W375–W379. [CrossRef] [PubMed]
38. Käll, L.; Krogh, A.; Sonnhammer, E.L.L. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **2004**, *338*, 1027–1036. [CrossRef] [PubMed]
39. Reynolds, S.M.; Kall, L.; Riffle, M.E.; Bilmes, J.A.; Noble, W.S. Transmembrane topology and signal peptide prediction using dynamic bayesian networks. *PLoS Comput. Biol.* **2008**, *4*, e1000213. [CrossRef] [PubMed]
40. Bendtsen, J.D.; Nielsen, H.; von Heijne, G.; Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **2004**, *340*, 783–795. [CrossRef] [PubMed]
41. Nielsen, H.; Krogh, A. Prediction of signal peptides and signal anchors by a hidden Markov model. In Proceedings of the International Conference on Intelligent Systems for Molecular Biology, Toronto, ON, Canada, 19–23 July 2008; Volume 6, pp. 122–130.
42. Maetschke, S.; Towsey, M.; Bodén, M. BLOMAP: An Encoding of Amino Acids which Improves Signal Peptide Cleavage Site Prediction. In Proceedings of the 3rd Asia-Pacific Bioinformatics Conference, Singapore, 17–21 January 2005; Imperial College Press: London, UK, 2005; pp. 141–150.
43. Murphy, L.R.; Wallqvist, A.; Levy, R.M. Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Eng.* **2000**, *13*, 149–152. [CrossRef] [PubMed]
44. Peterson, E.L.; Kondev, J.; Theriot, J.A.; Phillips, R. Reduced amino acid alphabets exhibit an improved sensitivity and selectivity in fold assignment. *Bioinformatics* **2009**, *25*, 1356–1362. [CrossRef] [PubMed]
45. von Heijne, G. A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res.* **1986**, *14*, 4683–4690. [CrossRef] [PubMed]
46. Cilingir, G.; Broschat, S.L.; Lau, A.O. ApicoAP: The first computational model for identifying apicoplast-targeted proteins in multiple species of Apicomplexa. *PLoS ONE* **2012**, *7*, e36598. [CrossRef] [PubMed]
47. Sperschneider, J.; Williams, A.H.; Hane, J.K.; Singh, K.B.; Taylor, J.M. Evaluation of Secretion Prediction Highlights Differing Approaches Needed for Oomycete and Fungal Effectors. *Front. Plant Sci.* **2015**, *6*, 1168. [CrossRef] [PubMed]

48. Yu, C.Y.; Li, X.X.; Yang, H.; Li, Y.H.; Xue, W.W.; Chen, Y.Z.; Tao, L.; Zhu, F. Assessing the performances of protein function prediction algorithms from the perspectives of identification accuracy and false discovery rate. *Int. J. Mol. Sci.* **2018**, *19*, 183. [CrossRef] [PubMed]
49. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152. [CrossRef] [PubMed]
50. Nielsen, H.; Engelbrecht, J.; Brunak, S.; von Heijne, G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **1997**, *10*, 1–6. [CrossRef] [PubMed]
51. Burdukiewicz, M.; Sobczyk, P.; Rödiger, S.; Duda-Madej, A.; Mackiewicz, P.; Kotulska, M. Amyloidogenic Motifs Revealed by N-Gram Analysis. *Sci. Rep.* **2017**, *7*, 12961. [CrossRef] [PubMed]
52. Dawson, D.M. *Size*; Academic Press: New York, NY, USA, 1972; pp. 1–38.
53. Fasman, G.D. *Proteins*, 3rd ed.; CRC Press: Cleveland, OH, USA, 1976; Volume 1.
54. Goldsack, D.E.; Chalifoux, R.C. Contribution of the free energy of mixing of hydrophobic side chains to the stability of the tertiary structure of proteins. *J. Theor. Biol.* **1973**, *39*, 645–651. [CrossRef]
55. Zimmerman, J.M.; Eliezer, N.; Simha, R. The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.* **1968**, *21*, 170–201. [CrossRef]
56. Cid, H.; Bunster, M.; Canales, M.; Gazitua, F. Hydrophobicity and structural classes in proteins. *Protein Eng.* **1992**, *5*, 373–375. [CrossRef] [PubMed]
57. Eisenberg, D. Three-dimensional structure of membrane and surface proteins. *Annu. Rev. Biochem.* **1984**, *53*, 595–623. [CrossRef] [PubMed]
58. Kyte, J.; Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982**, *157*, 105–132. [CrossRef]
59. Ponnuswamy, P.K.; Prabhakaran, M.; Manavalan, P. Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins. *Biochim. Biophys. Acta* **1980**, *623*, 301–316. [CrossRef]
60. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **1974**, *185*, 862–864. [CrossRef] [PubMed]
61. Radzicka, A.; Pedersen, L.; Wolfenden, R. Influences of solvent water on protein folding: Free energies of solvation of cis and trans peptides are nearly identical. *Biochemistry* **1988**, *27*, 4538–4541. [CrossRef] [PubMed]
62. Argos, P.; Rao, J.K.; Hargrave, P.A. Structural prediction of membrane-bound proteins. *Eur. J. Biochem.* **1982**, *128*, 565–575. [CrossRef] [PubMed]
63. Chou, P.Y.; Fasman, G.D. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol. Relat. Areas Mol. Biol.* **1978**, *47*, 45–148. [PubMed]
64. Prabhakaran, M. The distribution of physical, chemical and conformational properties in signal and nascent peptides. *Biochem. J.* **1990**, *269*, 691–696. [CrossRef] [PubMed]
65. Kawashima, S.; Pokarowski, P.; Pokarowska, M.; Kolinski, A.; Katayama, T.; Kanehisa, M. AAindex: Amino acid index database, progress report 2008. *Nucleic Acids Res.* **2008**, *36*, D202–D205. [CrossRef] [PubMed]
66. Rabiner, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **1989**, *77*, 257–286. [CrossRef]
67. Yu, S.Z. Hidden semi-Markov models. *Artif. Intell.* **2010**, *174*, 215–243. [CrossRef]
68. Koski, T. *Hidden Markov Models for Bioinformatics*; Computational Biology; Springer: Dordrecht, The Netherlands, 2001.
69. Durbin, R.; Eddy, S.R.; Krogh, A.; Mitchison, G. *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids*; Cambridge University Press: Cambridge, UK, 1998.