Michał Burdukiewicz1, Piotr Sobczyk2, Paweł Błażej1, Paweł Mackiewicz\*1,  
**1 University of Wrocław, Department of Genomics, Poland**  
**2 Wrocław University of Technology, Department of Mathematics, Poland**

\* pamac@smorfland.uni.wroc.pl

# Abstract

Signal peptides play an essential role in targeting substantial number of proteins to endomembrane system and their export outside the cell. Such proteins are of great importance in metabolism, maintenance of tissue structure, immune response and regulation of other organismal functions. The software for computional recognition of these peptides usually employs learning systems. To make correct predictions, these algorithms require big data sets consisting of signal peptides from different types of proteins and taxonomic groups. Therefore, they perform well in the identification of typical signal peptides but are ineffective in the prediction of atypical ones, for which it is not possible to obtain sufficiently large data sets for learning algorithms. Such class of peptides is present in proteins from parasites belonging to the phylum Apicomplexa characterized by a strongly AT-biased genomes and resulting from that specific amino acid composition of coded proteins. Their members are of great medically significance, especially *Plasmodium*, a malaria agent. That is why, we designed a new more flexible and universal probabilistic model for recognition of eukaryotic signal peptides, which includes knowledge about their organization, amino acid composition and variability. The proposed approach called signalHsmm is based on hidden semi-Markov models (HSMMs). It is able to recognize signal peptides from the malaria parasites and their relatives more accurately than popular programs, with the largest AUC = 0.94, MCC = 0.76 and the maximal sensitivity = 1.0. Simultaneously, it is still universal enough to provide prediction of other eukaryotic signal peptides on par with the best preforming predictors. Moreover, it proves to be very stable regardless of the learning data size. Therefore, our model does not need to be permanently retrained with the continuous expansion of sequence databases. The web-server of signalHsmm is available at <http://smorfland.uni.wroc.pl/signalhsmm>.

# Introduction

## Roles and features of signal peptides

Proteins of eukaryotes are encoded in nuclear genomes and are synthesized in ribosomes located in the cytosol or bounded by the endoplasmic reticulum. After translation, proteins are targeted to specific subcellular compartments or exported outside the cell. The proper localization of proteins is essential to perform their desired function. Information about the protein destination is included within the very protein in short stretches of amino acid residues called targeting or sorting signals. One kind of them are signal peptides, which are located at the N-terminus of proteins.

Signal peptides are responsible for targeting of proteins via the Sec61 translocation channel Rapoport (2007) to endomembrane system, which includes endoplasmic reticulum and Golgi apparatus. Such proteins can stay inside these compartments, can be inserted into cellular membranes or exported outside the cell. Proteins equipped with signal peptides play crucial role in metabolism (-galactosidase, pepsins) Hofmann and Schultz (1991), maintenance of tissue structure (collagen) Chan, Ho, and Cheah (2001), immune response (interferons, interleukins) L. Zhang, Leng, and Mixson (2005) and regulation of other organismal functions (prolactin, glucagon) Y. Huang, Wilkinson, and Willars (2010). Moreover, passing proteins through the endomembrane system is important for their correct folding and posttranslational modifications such as glycosylation and phosphorylation.

Despite the low sequence homology between signal peptides Ladunga (1999), some general architecture was proposed Izard and Kendall (1994; Voss, Schröder, and Fluhrer 2013) - Fig. [fig:sparch]. It is assumed that signal peptides start with a positively charged sequence of amino acid residues, called the n-region with the length of about 5-8 residues. They probably enforce a proper topology on a polypeptide during its translocation through the membrane based on the positive-inside rule G. von Heijne and Gavel (1988). The first region is followed by a stretch of hydrophobic amino acids (h-region) with the length of about 8-12 residues. It constitutes a core region of signal peptide and usually forms -helix. The third part of a signal peptide is a polar and uncharged c-region. It is usually 6 residues long and ends with a cleavage site, in which a signal peptidase cleaves the signal peptide, during or after translocation of the protein into the lumen of endoplasmic reticulum  Paetzel et al. (2002). The cleavage site is characterized by a variable amino acid composition. It typically contains small and neutral residues at -3 and -1 positions Palzkill et al. (1994). This site is, however, absent from some membrane proteins in which the first transmembrane domain acts both as a signal peptide and signal anchor Szczesna-Skorupa et al. (1988). The amino acid composition and the length of these regions vary between signal peptides, which influences the efficiency of protein secretion Hegde and Bernstein (2006).

![fig:](data:application/postscript;base64,) [fig:sparch]

On the other hand, signal peptides show a great variation and the description presented above (Fig. [fig:sparch]) refers to the most “typical” signal peptides. There are exceptionally long signal peptides, which fulfill more sophisticated roles Hiss and Schneider (2009). For example, the fragment of signal peptide from preprolactin takes part in the regulation of prolactin secretion, whereas signal peptides of MHC class I inhibit activity of NK cells. Signal peptides of viral origin are involved in the immune evasion or viral life cycle Kapp (2000) and the signal peptide from midkine, containing epitopes recognized by CD4+ T cells, contributes to tumor progression  Kerzerho et al. (2013).

The functional significance of these targeting signals makes that the prediction of signal peptide-containing proteins is an important step in the drug development L. Zhang, Leng, and Mixson (2005; Neto Ade et al. 2012; Moeller et al. 2010). The signal peptides can be potential drug targets, especially for malaria parasites from the genus *Plasmodium* belonging to the phylum Apicomplexa. Their signal peptides not only direct proteins into endomembrane system or outside the cell but also into a unique organellum specific only for *Apicomplexa*, called apicoplast Foth and McFadden (2003; Lim and McFadden 2010; McFadden 2011; Heiny et al. 2014). It is a reduced four-membrane plastid, which lost photosynthetic function but play an important role in synthesis of fatty acids and lipids Lim and McFadden (2010; McFadden 2011; Mazumdar et al. 2006). Therefore, anti-malarial drugs disturbing pathways targeting proteins into the apicoplast should be safe for human (host) and harmful for the parasite Fichera and Roos (1997; Ralph, D’Ombrain, and McFadden 2001; Gornicki 2003; Garcia-Estrada et al. 2010). Other potential apicomplexan drug target, the food vacuole, also imports proteins tagged with the signal peptide Egan (2002). A model of *Plasmodium* signal peptide would aid in the design of appropriate medicines. However, the targeting signals of *Plasmodium* proteins show deviated composition in comparison to typical signals because of heavy adenine-thymine bias of parasitic genomes Tonkin, Kalanon, and McFadden (2008). Therefore, the current programs dedicated to recognition of general eukaryotic signal peptides are not appropriate.

## Software predicting signal peptides

Since experimental methods determining the subcellular localization of proteins and identifying signal peptide are time-consuming and laborious, different computational approaches predicting targeting signals were developed. Among them, signal peptides became the subject of several computational programs to their prediction. Many software incorporates ’black-box’ models, such as: neural networks Petersen et al. (2011), support vector machines S.-W. Zhang et al. (2014), Bayesian networks Zheng et al. (2012) or k-nearest neighbours Shen and Chou (2007). However, these models do not provide direct biological information about organization of signal peptides and are not able to predict properly atypical signal peptides. Although there are programs that do not share the innate flaws of ’black-box’ models, they also demand an improvement. Some of them are based on position matrices or their variants S.-W. Zhang et al. (2014; Hiller et al. 2004). Others (Phobius, Philius and SignalP 3.0) use hidden Markov models (HMMs) Käll, Krogh, and Sonnhammer (2004; Reynolds et al. 2008; Bendtsen et al. 2004), which try to reflect structure of signal peptides regions in their limited probabilistic frameworks. The used HMMs, however, imply a geometric distribution for duration of regions length. We studied the distribution for regions from the first work utilizing HMMs in prediction of signal peptides H. Nielsen and Krogh (1998) and found that the length distribution for every region was not geometric (Fig. [fig:reglen]).

Majority of the signal peptide predicting software uses the orthogonal encoding of amino acids, in which a vector of 20 digits represents every amino acid. This method of encoding, however, does not take into account relationships between amino acids and differences in their physicochemical properties. This is disadvantage of such signal peptide’s models because their regions are in fact characterized by specific features of amino acid residues and not by the simple occurrence of particular amino acids. In addition to this, such sparse encoding requires large data sets, which hinders their management and analysis Lin, May, and Taylor (2002).

Moreover, the existing algorithms require a large number of sequences to be successfully learned. Although their learning sets are constructed to be as diverse as possible, the most frequent sequences dominate creation of decision rules. As a result, such predictiors usually accurately identify most common signal peptides, but are ineffective in more atypical cases, which are less represented in training data set. The commonly used rigid scheme of signal peptide’s organization (Fig. [fig:sparch]) does not characterize extremely long or short peptides, which constitute substantial fraction of all signal peptides.

Therefore, we elaborated a new approach called signalHsmm. To enable the prediction of signal peptides with atypical amino acid composition, as in *Plasmodiidae*, we use grouping of amino acids into physicochemical groups. Instead of looking for patterns consisting of specific residues, we focus on more general properties that are necessary for functional signal peptide. This approach is supported by the recent advancements in the proteomics suggesting that such simplification of amino acid alphabet (by reducing the number of letters) may lead to better fold recognition Murphy, Wallqvist, and Levy (2000; Peterson et al. 2009). Considering that one of key features of signal peptide is its secondary structure (-helix), the universal model may utilize shorter amino acid alphabet, where several similar amino acids are unified into a single group.

In addition the the alphabet reduction, our algorithm is based on hidden semi-Markov models. This flexible probabilistic framework does not assume the geometric distribution of the region length, but rather learns such distribution from training data set. Hence, signalHsmm utilizing hidden semi-Markov models and a reduced amino acid alphabet, should be more general than its counterparts.

![fig:](data:application/postscript;base64,) [fig:reglen]

# Materials and Methods

## Overview

Since the functionality of signal peptides depends on the physicochemical properties of residues in a given region, we clustered amino acids into several groups based on their features. The pre-processed sequences were further analyzed by an heuristic algorithm, which determines borders between three characteristic regions in signal peptides, which is an enhanced version of heuristic algorithm employed in signalP 2.0 H. Nielsen and Krogh (1998). We refined some criteria in recognition of the regions to attune the algorithm to less typical signal peptides. Next, two models were trained to recognize proteins with and without a signal peptide. The first one was a hidden semi-Markov model, in which each of three signal peptide’s regions was represented by a different hidden state. The additional fourth hidden state represented a 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 signal peptide cleavage sites. The second model was a simple probabilistic approach in which no association between amino acids was assumed and probability of amino acids groups occurrence was determined by their frequencies in mature proteins.

## Data selection

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

The main testing set consisting of proteins belonging to the organisms from *Plasmodiidae* family. The positive set contained 102 sequences with a putative signal peptide having annotated start and cleavage site. The corresponding negative set comprised 358 sequences without any signal peptide information. The other testing set consists of 127 eukaryotic proteins with signal peptide included in the UniProt database after 2010 year.

## Homology reduction of studied sequence sets

To reduce the set according to homology of collected protein sequences, we filtered them using cd-hit Fu et al. (2012). The homology reduction were subjected sequences of signal peptides and the first 70 amino acid residues in the case of proteins without the peptide, as proposed by H. Nielsen et al. (1997). We prepared two reduced learning data sets with data deposited till 2010 and 1987 year by removing the homology on 50%-similarity threshold with word length 2. After this procedure, these sets contained 748 and 132 sequences with signal peptide, respectively. The testing set was filtered in the same manner. After that, these sets were reduced to 51 and 211 sequences with and without signal peptide, respectively.

## Clustering of amino acids into groups

To reduce alphabet of amino acids, we clustered them into several physico-chemical groups. It is a different approach in comparison to BLOMAP Maetschke, Towsey, and Bodén (2005), which also uses a reduced alphabet of amino acids, but based on substitution matrices. We grouped amino acids using four properties relevant for the architecture of signal peptide: their hydrophobicity, tendency to occurring in -helices, polarity and size. The high hydrophobicity is a good determinant of the h-region, whose -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 Palzkill et al. (1994).

[ht]

ll Property name & Amino acid scale  
Size & Size Dawson (1972)  
Size & Molecular weight Fasman (1976)  
Size & Residue volume Goldsack and Chalifoux (1973)  
Size & Bulkiness Zimmerman, Eliezer, and Simha (1968)  
Hydrophobicity & Normalized hydrophobicity scales for -proteins Cid et al. (1992)  
Hydrophobicity & Consensus normalized hydrophobicity scale Eisenberg (1984)  
Hydrophobicity & Hydropathy index Kyte and Doolittle (1982)  
Hydrophobicity & Surrounding hydrophobicity in -helix Ponnuswamy, Prabhakaran, and Manavalan (1980)  
Polarity & Polarity Grantham (1974)  
Polarity & Mean polarity Radzicka, Pedersen, and Wolfenden (1988)  
Occurrence in -helices & Signal sequence helical potential Argos, Rao, and Hargrave (1982)  
Occurrence in -helices & Normalized frequency of N-terminal helix P. Y. Chou and Fasman (1978)  
Occurrence in -helices & Relative frequency in -helix Prabhakaran (1990)

[tab:aaprop]

We considered in total 13 amino acid scales present in AAIndex database Kawashima et al. (2008) (Tab. [tab:aaprop]). 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 group 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 our algorithm on every encoding. We created balanced data sets by subsampling proteins without a signal peptide to equal the number of proteins with a signal peptide. The cross-validation was repeated 60 times to ensure that every protein without signal peptide was included in the learning set with the probability higher than 0.5.

## Hidden semi-Markov model

Our algorithm is based on hidden semi-Markov model (HSMM), which is an extension of hidden Markov model (HMM) Rabiner (1989; Yu 2010; Koski 2001). The HMM consists of two stochastic processes. The first is a discrete Markov chain on the set of hidden states , where means a step of this process and means the total duration of the process corresponding to the length of signal peptide. The hidden states represent particular signal peptide regions and are “the cause” of the observations, which are amino acid residues in analyzed sequences. In a subsequent step , the hidden state might change to another according to a transition matrix , where means a probability of being in a state on condition that in the previous step was a state . The second process is an observation process defined on the set of possible observations . They are assumed to occur independently but conditionally on the hidden states that emits these observations. The distribution of observations are given by a matrix , where means a probability of emission of observation on condition that a hidden state was . The main goal of signalHsmm is to find the most probable signal peptide’s regions boundaries for a given sequences. This is achieved with 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. Durbin et al. (1998) showed how to extend it for different distributions without significant increase in computational complexity. Similar ideas were used for signal peptide recognition, for example by Käll, Krogh, and Sonnhammer (2004). However, it is still not flexible enough because the empirical regional length distributions (see Fig. [fig:reglen]) are difficult to capture in this way.

= [draw,shape=circle, top color=green!50!white!70, bottom color=green!50!white!70 ,minimum size=4em] = [draw,shape=circle, top color=red!50!white!70, bottom color=red!50!white!70 ,minimum size=2.5em] =[shape=rectangle, top color=white, bottom color=white ,minimum size=3pt,inner sep=0pt] =[shape=rectangle, top color=white, bottom color=white ,minimum size=3pt,inner sep=0pt] = [draw=black, color=black!70!white!50, line width=1.5mm, -latex’] = [draw=black, color=red!30!white!70, line width=1.5mm, -latex’] =[text=black,above, bottom color=white, top color=blue!50!black!70 ] =[text=black,below, bottom color=red!50!white!70, top color=red!50!white!70 ] = [color=green!50!black!70, line width=1.5mm]

[>=latex’]

at (1+2,0) (block1) ; at (-.5+2,-2) (komp1) ; at (1+2,-2) (komp2) ; at (2.5+2,-2) (komp3) ; (block1) – (komp1); (block1) – (komp2); (block1) – (komp3); (block1.east) – +(2,0) node [branch,midway,yshift=0.3cm,xshift=-0.2cm,color=black] **transition**;

at (4.5+2,0) (block2) ...; (block2.east) – +(2,0) node [branch,midway,yshift=0.3cm,xshift=-0.2cm,color=black] **transition**; at (4.5+2,-2) (komp4) ; (block2) – (komp4) ;

at (8+2,0) (block3) ; at (6.5+2,-2) (komp5) ; at (8+2,-2) (komp6) ; at (9.5+2,-2) (komp7) ; (block3) – (komp5); (block3) – (komp6); (block3) – (komp7);

(1,-1.8) – (5,-1.8) node [branch2, midway, yshift=-1cm,color=black] **Duration of length** ;

(8,-1.8) – (12,-1.8) node [branch2, midway, yshift=-1cm,color=black] **Duration of length** ;

(1,-2.6) – (12,-2.6) node [branch2, midway, yshift=-1.3cm,color=black] **Total duration,** ;

[fig:hsmm]

In our approach, we used a modification of HMM called hidden semi-Markov model (HSMM) Yu (2010). It extends the HMM by assuming a duration distribution for a given hidden state (Fig. [fig:hsmm]). Then, the model includes additionally probabilities of duration in hidden states:

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

Almost all entries in the transition matrix are zeros because regions represented by hidden states are sequential. Possible transitions between them are depicted as arrows in Fig. [fig:ngramext]. Probabilities of observations for the hidden states and hidden states durations were estimated from training data. The advantage of HSMM model results not only from its better performance but also form its straightforwardness and flexibility.

![fig:](data:application/postscript;base64,) [fig:ngramext]

# Results and discussion

## Performance of signalHsmm algorithm

![fig:](data:application/postscript;base64,) [fig:cvres]

[ht]

.5

ll 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

[tab:best]

.5

ll 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

[tab:worst]

![fig:](data:application/postscript;base64,) [fig:enccomp]

To evaluate efficiency of our algorithm, we calculated four performance measures after cross-validation procedure for all amino acid encodings: specificity, sensitivity, Matthew’s Correlation Coefficient ( coefficient) and Area Under the Curve (AUC). The measures were characterized by very small variance (see for example Tab. [tab:perfmeas]), which indicates the credibility of the applied 5-fold cross-validation with 60 repetitions. All encodings of amino acids showed very good and quite narrow range of AUC (0.93 – 0.97) and specificity (0.92 – 0.96). However, sensitivity characterized by much wider variation and ranged from 0.66 to 0.94 (Fig. [fig:cvres]). 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 (Tab. [tab:best]).

[ht] [tab:perfmeas]

lrr Measure & Mean & SD  
AUC & 0.9682 & 0.0023  
Sensitivity & 0.9407 & 0.0008  
Specificity & 0.9272 & 0.0050  
MCC & 0.8681 & 0.0049

## Comparison of amino acid encodings

We examined in detail composition of encodings and properties of their amino acids with the best sensitivity and the best specificity (Tab. [tab:best], Tab. [tab:worst] and Fig. [fig:enccomp]). In both cases, the group 1 tends to contain generally average-sized polar amino acids. This group in the best sensitivity encoding is more uniform because it includes all charged amino acids, both acidic and basic (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 h-region and provide very good distinction between regions of signal peptide (Fig. [fig:enccomp]). In the best specificity encoding, where polar and charged character of the group 1 is not so explicit, the difference in its distribution between the regions is also less visible.

The amino acids belonging to the group 2 show generally a quite low probability of occurrence in -helix. The best sensitivity encoding comprises two types amino acids: all three hydroxylated residues as well as aliphatic glycine and proline known to break -helices. The best specificity encoding lacks tyrosine but includes in addition aspartic acid and its amide, which increase the polar character of this group. Despite these differences, the occurrence of the group 2 from both encodings is very similar in signal peptide’s regions (Fig. [fig:enccomp]). 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 the both amino acid classifications (Fig. [fig:enccomp]).

The fourth group 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 -helices. This very unique composition seems to be the most typical to the c-region of signal peptide. In contrast, the group 4 in the best specificity encoding contains large aromatic amino acids: phenylalanine, tryptophan and tyrosine without special preference to signal peptide’s regions.

The encodings of amino acid plays crucial role in the recognition of signal peptide but does not affect in such extent identification of proteins without signal peptides. The change in specificity for different encodings is seven times smaller than for sensitivity (Fig. [fig:cvres]). It results from more uniform distribution of different residues in the part of mature protein than in signal peptide’s regions.

## Benchmark tests

To provide the fair comparison of our algorithm with previous software, we trained our model on 2311 signal peptide-containing sequences deposited in Uniprot until 2010 year (the iteration of signalHsmm called signalHsmm-2010). The set should correspond to data used to train SignalP 4.1, the newest classifier present in the benchmark. In addition to this, we prepared the smaller data set, covering only 336 sequences collected till 1987 year, just after the first method predicting signal peptide was published G. von Heijne (1986). The signalHsmm-1987 iteration had to extract the signal peptide model from data set more limited than training sets of any classifier included in the benchmark.

Together with signalHsmm, we evaluated several signal peptide predicting algorithms in recognition of atypical signal peptides from malaria parasites: SignalP 4.1, PrediSi, Phobius and Philius (see Tab. [tab:bench2010plas] for the most common performance measures, and Supplemental Table [tab:bench2010plasfull] for 23 performance measures). The older version of SignalP 3.0, was also incorporated in the analysis, because it is often chosen over its newer counterpart in the analysis of sequences belonging to Apicomplexa and other taxa due to the larger sensitivity  Cilingir, Broschat, and Lau (2012; Sperschneider et al. 2015). 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 [tab:bench2010plasfull]). PrediSi received the best specificity but at the expense of significantly reduced sensitivity.

We trained several iterations of signalHsmm described above to check improvements introduced to our software, i.e. a new probabilistic model (hidden semi-Markov model) and a simplified alphabet of amino acids. The latter was compared with the version trained on raw amino acids sequences, denoted as ’raw aa’ in Tab. [tab:bench2010plas]. The performance of this iteration was mostly worse than 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 [tab:bench2010plasfull]). This result confirms that unique function of signal peptide does not depend on specific amino acids, but on more general features and our simpler model is able to recognize the unique architecture of signal peptide more accurately. The advantage of hidden semi-Markov model over normal Markov model can be seen through the comparison of the signalHsmm with signal peptide predictors utilizing HMM: Phobius, Philius and SignalP 3.0 (HMM) - Tab. [tab:bench2010plas]. For 17 of 23 measures including AUC, MCC and sensitivity, our algorithm performed best (Tab. [tab:bench2010plas], Supplemental Table [tab:bench2010plasfull]).

To check susceptibility of our model to overfitting, we trained it on data sets with 50%-sequence similarity 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 homology reduction were comparable. What is more, the former was slightly better in 21 measures than the model based on all sequences.

[ht] [tab:bench2010plas]

rllll & Sensitivity & Specificity & MCC & AUC  
signalP 4.1 (no tm) Petersen et al. (2011) & 0.8235 & 0.9100 & 0.6872 & 0.8667  
signalP 4.1 (tm) Petersen et al. (2011) & 0.6471 & 0.9431 & 0.6196 & 0.7951  
signalP 3.0 (NN) Bendtsen et al. (2004) & 0.8824 & 0.9052 & 0.7220 & 0.8938  
signalP 3.0 (HMM) Bendtsen et al. (2004) & 0.6275 & 0.9194 & 0.5553 & 0.7734  
PrediSi Hiller et al. (2004) & 0.3333 & **0.9573** & 0.3849 & 0.6453  
Philius Reynolds et al. (2008) & 0.6078 & 0.9336 & 0.5684 & 0.7707  
Phobius Käll, Krogh, and Sonnhammer (2004) & 0.6471 & 0.9289 & 0.5895 & 0.7880  
signalHsmm-2010 & 0.9804 & 0.8720 & 0.7409 & 0.9262  
signalHsmm-2010 (hom. 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 (hom. 50%) & 0.9412 & 0.8768 & 0.7194 & 0.9090  
signalHsmm-1987 (raw aa) & 0.7647 & 0.9052 & 0.6350 & 0.8350

The overall simplicity of our approach does not hinder its capabilities of recognizing signal peptides from other organisms. We benchmarked signalHsmm iterations and other software on the set of 127 eukaryotic proteins with signal peptide added after year 2010 to UniProt and randomly chosen 127 proteins without signal peptide. Their homology in the testing set was also reduced as described above. SignalHsmm-2010 performed comparably to SignalP 4.1 (see Supplemental Table [tab:bench2010full] for all performance measures). Its AUC was 0.94 in comparison to 0.95 and 0.96 of two SignalP 4.1 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.

## Specific composition of *Plasmodiidae* signal peptides

![fig:](data:application/postscript;base64,) [fig:aa]

![fig:](data:application/postscript;base64,) [fig:PCA]

Since *Plasmodium* genomes are characterized by a large excess of adenine and thymine Tonkin, Kalanon, and McFadden (2008), this bias strongly influences amino acid composition of coded proteins including signal peptides (Fig. [fig:aa]). These targeting signals differ significantly between *Plasmodiidae* and other taxa in composition of 13 amino acids (Wilcoxon test, p corrected by the Benjamini-Hochberg method in the multiple testing). The *Plasmodiidae* signal peptides are especially abundant in amino acid coded by codons rich in A and T, such as: phenylalanine (TTY), isoleucine (ATH), lysine (AAR) and asparagine (AAY), whereas are poor in amino acids coded by GC-rich codons: alanine (GCN), glycine (GGN) and proline (CCN). Leucine is coded by set of codons with mixed composition (TTR, CTN) but also discriminate the two sets of signal peptides.

As a result of this, signal peptides from *Plasmodiidae* separate from other signal peptides according to raw amino acid composition in Principal Component Analysis (Fig. [fig:PCA]A). Mature proteins of *Plasmodiidae* are also shifted from the other mature proteins. Therefore, algorithms that consider only particular amino acids residues may do not have decision rules appropriate for such composition and consequently fail to identify such peptides. Interestingly, the amino acid encoding employed by signalHsmm reduces this difference and unifies all signal peptides into one set (Fig. [fig:PCA]B). Similarly, when the reduced amino acid alphabet is considered, mature proteins are also inseparable. It should be emphasized that the degeneration of amino acids utilized by our algorithm does not weaken the difference between signal peptides and mature proteins, which still create distinguishable groups. These analyzes indicate that the applied amino acid encoding enables to capture a general composition of many signal peptides keeping simultaneously their difference from mature proteins.

# Conclusions

We proposed a novel solution to the problem of predicting signal peptides, which appeared very efficient in recognition of atypical signal peptides from proteins of *Plasmodiidae* although the program was trained on data coming from all eukaryotes. It indicates that our algorithm is able to describe common features of all signal peptides basing on the classical division of signal peptide into three regions. Our software is not limited to very specific taxonomic group, but is able to compete with state-of-art algorithms in predicting signal peptides of other organisms.

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

The existing software detecting signal peptides does not usually reveal decision rules responsible for this prediction. Our algorithm is the first step to explicitly show features of signal peptides 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 regions of signal peptide. The 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 -helix breakers: glycine and proline.

The flexibility and efficiency in recovery information make our model unique among similar software. SignalHsmm models properly very specific signal peptides belonging to narrow taxonomic groups which are poorly represented in databases and can effectively extract information from very small data sets. Our approach may lead in future to development of new predictors specialized in recognition of atypical signals targeting sequences to subcellular compartments.

# Availability and implementation

The signalHsmm prediction web-server is available at: <http://smorfland.uni.wroc.pl/signalhsmm>. SignalHsmm is implemented as an R package available at: <http://cran.r-project.org/web/packages/signalHsmm>. Stand-alone version offers prediction and tools to build, train and test novel signal peptide models.

# Supporting Information

## Table

**Benchmark results - *Plasmodiidae*.** Performance measures for benchmark test of signal peptide predictors using sequences belonging to *Plasmodiidae*.

## Table

**Benchmark results - Eukaryots** Performance measures for benchmark test of signal peptide predictors using all eukaryotic sequences.

# Acknowledgments

# Author Contributions

MB and PM designed the study. All authors participated in the design of the study. PB and PS created the probabilistic model. MB and PS implemented the model and wrote the software. MB and PM wrote the manuscript. All authors critically revised the manuscript.

# References

Argos, P., J. K. Rao, and P. A. Hargrave. 1982. “Structural Prediction of Membrane-Bound Proteins.” *European Journal of Biochemistry* 128 (2-3): 565–75.

Bendtsen, J. D., H. Nielsen, G. von Heijne, and S. Brunak. 2004. “Improved Prediction of Signal Peptides: SignalP 3.0.” *J Mol Biol* 340 (4): 783–95.

Chan, Danny, Matthew S. P. Ho, and Kathryn S. E. Cheah. 2001. “Aberrant Signal Peptide Cleavage of Collagen X in Schmid Metaphyseal Chondrodysplasia. Implications for the Molecular Basis of the Disease.” *Journal of Biological Chemistry* 276 (11): 7992–7997.

Chou, P. Y., and G. D. Fasman. 1978. “Prediction of the Secondary Structure of Proteins from Their Amino Acid Sequence.” *Advances in Enzymology and Related Areas of Molecular Biology* 47: 45–148.

Cid, H., M. Bunster, M. Canales, and F. Gazitua. 1992. “Hydrophobicity and Structural Classes in Proteins.” *Protein Eng* 5 (5): 373–5.

Cilingir, G., S. L. Broschat, and A. O. Lau. 2012. “ApicoAP: the First Computational Model for Identifying Apicoplast-Targeted Proteins in Multiple Species of Apicomplexa.” *PLoS ONE* 7 (5): e36598.

Dawson, D. M. 1972. *Size*. Edited by D.J.H. Brock and O. Mayo. New York: Academic Press.

Durbin, Richard, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. 1998. *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Egan, T. J. 2002. “Discovering Antimalarials: a New Strategy.” *Chem Biol* 9 (8): 852–3.

Eisenberg, D. 1984. “Three-Dimensional Structure of Membrane and Surface Proteins.” *Annu Rev Biochem* 53: 595–623.

Fasman, G. D. 1976. *Proteins*. 3rd ed. Vol. 1. Cleveland: CRC Press.

Fichera, M. E., and D. S. Roos. 1997. “A Plastid Organelle as a Drug Target in Apicomplexan Parasites.” *Nature* 390 (6658): 407–9.

Foth, B. J., and G. I. McFadden. 2003. “The Apicoplast: a Plastid in Plasmodium Falciparum and Other Apicomplexan Parasites.” *Int Rev Cytol* 224: 57–110.

Fu, L., B. Niu, Z. Zhu, S. Wu, and W. Li. 2012. “CD-HIT: Accelerated for Clustering the Next-Generation Sequencing Data.” *Bioinformatics* 28 (23): 3150–2.

Garcia-Estrada, C., C. F. Prada, C. Fernandez-Rubio, F. Rojo-Vazquez, and R. Balana-Fouce. 2010. “DNA Topoisomerases in Apicomplexan Parasites: Promising Targets for Drug Discovery.” *Proc Biol Sci* 277 (1689): 1777–87.

Goldsack, D. E., and R. C. Chalifoux. 1973. “Contribution of the Free Energy of Mixing of Hydrophobic Side Chains to the Stability of the Tertiary Structure of Proteins.” *J Theor Biol* 39 (3): 645–51.

Gornicki, P. 2003. “Apicoplast Fatty Acid Biosynthesis as a Target for Medical Intervention in Apicomplexan Parasites.” *Int J Parasitol* 33 (9): 885–96.

Grantham, R. 1974. “Amino Acid Difference Formula to Help Explain Protein Evolution.” *Science* 185 (4154): 862–4.

Hegde, Ramanujan S, and Harris D Bernstein. 2006. “The Surprising Complexity of Signal Sequences.” *Trends in Biochemical Sciences* 31: 563–571.

Heijne, G. von. 1986. “A New Method for Predicting Signal Sequence Cleavage Sites.” *Nucleic Acids Res* 14 (11): 4683–90.

Heijne, G. von, and Y. Gavel. 1988. “Topogenic Signals in Integral Membrane Proteins.” *European Journal of Biochemistry* 174 (4): 671–8.

Heiny, Sabrina R., Sabine Pautz, Mario Recker, and Jude M. Przyborski. 2014. “Protein Traffic to the Plasmodium Falciparum Apicoplast: Evidence for a Sorting Branch Point at the Golgi.” *Traffic* 15 (12) (Dec): 1290–1304. doi:[10.1111/tra.12226](http://dx.doi.org/10.1111/tra.12226). <http://dx.doi.org/10.1111/tra.12226>.

Hiller, Karsten, Andreas Grote, Maurice Scheer, Richard Münch, and Dieter Jahn. 2004. “PrediSi: Prediction of Signal Peptides and Their Cleavage Positions.” *Nucleic Acids Research* 32: W375–W379.

Hiss, J. A., and G. Schneider. 2009. “Architecture, Function and Prediction of Long Signal Peptides.” *Brief Bioinform* 10 (5): 569–78.

Hofmann, Kathryn J., and Loren D. Schultz. 1991. “Mutations of the Α-Galactosidase Signal Peptide Which Greatly Enhance Secretion of Heterologous Proteins by Yeast.” *Gene* 101 (1): 105–111.

Huang, Y., G. F. Wilkinson, and G. B. Willars. 2010. “Role of the Signal Peptide in the Synthesis and Processing of the Glucagon-Like Peptide-1 Receptor.” *British Journal of Pharmacology* 159 (1): 237–251.

Izard, Jennifer W., and Debra A. Kendall. 1994. “Signal Peptides: Exquisitely Designed Transport Promoters.” *Molecular Microbiology* 13 (5): 765–773.

Käll, Lukas, Anders Krogh, and Erik L L Sonnhammer. 2004. “A Combined Transmembrane Topology and Signal Peptide Prediction Method.” *Journal of Molecular Biology* 338: 1027–1036.

Kapp, M. K.; Dobberstein, K.; Schrempf S.; Lemberg. 2000. “Post-Targeting Functions of Signal Peptides.” Edited by R Zimmermann. *Protein Transport into the Endoplasmic Reticulum, Madame Curie Bioscience Database*.

Kawashima, S., P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa. 2008. “AAindex: Amino Acid Index Database, Progress Report 2008.” *Nucleic Acids Res* 36 (Database issue): D202–5.

Kerzerho, J., A. Schneider, E. Favry, F. A. Castelli, and B. Maillere. 2013. “The Signal Peptide of the Tumor-Shared Antigen Midkine Hosts CD4+ T Cell Epitopes.” *J Biol Chem* 288 (19): 13370–7.

Koski, T. 2001. *Hidden Markov Models for Bioinformatics*. Computational Biology. Springer Netherlands. <https://books.google.pl/books?id=-VDqvCaYv4MC>.

Kyte, J., and R. F. Doolittle. 1982. “A Simple Method for Displaying the Hydropathic Character of a Protein.” *J Mol Biol* 157 (1): 105–32.

Ladunga, I. 1999. “PHYSEAN: PHYsical SEquence ANalysis for the Identification of Protein Domains on the Basis of Physical and Chemical Properties of Amino Acids.” *Bioinformatics* 15 (12): 1028–38.

Lim, Liting, and Geoffrey Ian McFadden. 2010. “The Evolution, Metabolism and Functions of the Apicoplast.” *Philos Trans R Soc Lond B Biol Sci* 365 (1541) (Mar): 749–763. doi:[10.1098/rstb.2009.0273](http://dx.doi.org/10.1098/rstb.2009.0273). <http://dx.doi.org/10.1098/rstb.2009.0273>.

Lin, K., A. C. May, and W. R. Taylor. 2002. “Amino Acid Encoding Schemes from Protein Structure Alignments: Multi-Dimensional Vectors to Describe Residue Types.” *J Theor Biol* 216 (3): 361–65.

Maetschke, Stefan, Michael Towsey, and Mikael Bodén. 2005. “BLOMAP: An Encoding of Amino Acids Which Improves Signal Peptide Cleavage Site Prediction.” In *In Chen Y., Wong L: Proc. 3 Rd AsiaPacific Bioinformatics Conference, Imperial*, 141–150. College Press.

Mazumdar, Jolly, Emma H Wilson, Kate Masek, Christopher A Hunter, and Boris Striepen. 2006. “Apicoplast Fatty Acid Synthesis Is Essential for Organelle Biogenesis and Parasite Survival in Toxoplasma Gondii.” *Proc Natl Acad Sci U S A* 103 (35) (Aug): 13192–13197. doi:[10.1073/pnas.0603391103](http://dx.doi.org/10.1073/pnas.0603391103). <http://dx.doi.org/10.1073/pnas.0603391103>.

McFadden, Geoffrey Ian. 2011. “The Apicoplast.” *Protoplasma* 248: 641–650.

Moeller, L., R. Taylor-Vokes, S. Fox, Q. Gan, L. Johnson, and K. Wang. 2010. “Wet-Milling Transgenic Maize Seed for Fraction Enrichment of Recombinant Subunit Vaccine.” *Biotechnol Prog* 26 (2): 458–65.

Murphy, L. R., A. Wallqvist, and R. M. Levy. 2000. “Simplified Amino Acid Alphabets for Protein Fold Recognition and Implications for Folding.” *Protein Eng* 13 (3): 149–52.

Neto Ade, M., D. A. Alvarenga, A. M. Rezende, S. S. Resende, S. Ribeiro Rde, C. J. Fontes, L. H. Carvalho, and C. F. de Brito. 2012. “Improving N-Terminal Protein Annotation of Plasmodium Species Based on Signal Peptide Prediction of Orthologous Proteins.” *Malar J* 11: 375.

Nielsen, H, and A Krogh. 1998. “Prediction of Signal Peptides and Signal Anchors by a Hidden Markov Model.” *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* 6: 122–130.

Nielsen, H, J Engelbrecht, S Brunak, and G von Heijne. 1997. “Identification of Prokaryotic and Eukaryotic Signal Peptides and Prediction of Their Cleavage Sites.” *Protein Engineering* 10: 1–6.

Paetzel, M., A. Karla, N. C. Strynadka, and R. E. Dalbey. 2002. “Signal Peptidases.” *Chem Rev* 102 (12): 4549–80.

Palzkill, T, Q Q Le, A Wong, and D Botstein. 1994. “Selection of Functional Signal Peptide Cleavage Sites from a Library of Random Sequences.” *Journal of Bacteriology* 176 (3): 563–568.

Petersen, Thomas Nordahl, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. 2011. “SignalP 4.0: Discriminating Signal Peptides from Transmembrane Regions.” *Nature Methods* 8: 785–786.

Peterson, E. L., J. Kondev, J. A. Theriot, and R. Phillips. 2009. “Reduced Amino Acid Alphabets Exhibit an Improved Sensitivity and Selectivity in Fold Assignment.” *Bioinformatics* 25 (11): 1356–62.

Ponnuswamy, P. K., M. Prabhakaran, and P. Manavalan. 1980. “Hydrophobic Packing and Spatial Arrangement of Amino Acid Residues in Globular Proteins.” *Biochim Biophys Acta* 623 (2): 301–16.

Prabhakaran, M. 1990. “The Distribution of Physical, Chemical and Conformational Properties in Signal and Nascent Peptides.” *Biochem J* 269 (3): 691–6.

Rabiner, Lawrence R. 1989. “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.” *Proceedings of the IEEE* 77 (2): 257–286.

Radzicka, A., L. Pedersen, and R. Wolfenden. 1988. “Influences of Solvent Water on Protein Folding: Free Energies of Solvation of Cis and Trans Peptides Are Nearly Identical.” *Biochemistry* 27 (12): 4538–41.

Ralph, S. A., D’OmbrainM. C., and G. I. McFadden. 2001. “The Apicoplast as an Antimalarial Drug Target.” *Drug Resist Updat.* 4: 145–151.

Rapoport, T. A. 2007. “Protein Translocation Across the Eukaryotic Endoplasmic Reticulum and Bacterial Plasma Membranes.” *Nature* 450 (7170): 663–9.

Reynolds, S. M., L. Kall, M. E. Riffle, J. A. Bilmes, and W. S. Noble. 2008. “Transmembrane Topology and Signal Peptide Prediction Using Dynamic Bayesian Networks.” *PLoS Comput Biol* 4 (11): e1000213.

Shen, Hong-Bin, and Kuo-Chen Chou. 2007. “Signal-3L: A 3-Layer Approach for Predicting Signal Peptides.” *Biochemical and Biophysical Research Communications* 363: 297–303.

Sperschneider, Jana, Angela H. Williams, James K. Hane, Karam B. Singh, and Jennifer M. Taylor. 2015. “Evaluation of Secretion Prediction Highlights Differing Approaches Needed for Oomycete and Fungal Effectors.” *Frontiers in Plant Science* 6 (December). doi:[10.3389/fpls.2015.01168](http://dx.doi.org/10.3389/fpls.2015.01168). <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4688413/>.

Szczesna-Skorupa, E., N. Browne, D. Mead, and B. Kemper. 1988. “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 U S A* 85 (3): 738–742.

Tonkin, Christopher J., Ming Kalanon, and Geoffrey I. McFadden. 2008. “Protein Targeting to the Malaria Parasite Plastid.” *Traffic* 9 (2) (Feb): 166–175. doi:[10.1111/j.1600-0854.2007.00660.x](http://dx.doi.org/10.1111/j.1600-0854.2007.00660.x). <http://dx.doi.org/10.1111/j.1600-0854.2007.00660.x>.

Voss, Matthias, Bernd Schröder, and Regina Fluhrer. 2013. “Mechanism, Specificity, and Physiology of Signal Peptide Peptidase (SPP) and SPP-Like Proteases.” *Biochimica Et Biophysica Acta* 1828: 2828–2839.

Yu, Shun-Zheng. 2010. “Hidden Semi-Markov Models.” *Artificial Intelligence* 174 (2): 215–243. doi:[http://dx.doi.org/10.1016/j.artint.2009.11.011](http://dx.doi.org/http://dx.doi.org/10.1016/j.artint.2009.11.011). <http://www.sciencedirect.com/science/article/pii/S0004370209001416>.

Zhang, L., Q. Leng, and A. J. Mixson. 2005. “Alteration in the IL-2 Signal Peptide Affects Secretion of Proteins in Vitro and in Vivo.” *J Gene Med* 7 (3): 354–65.

Zhang, Shao-Wu, Ting-He Zhang, Jun-Nan Zhang, and Yufei Huang. 2014. “Prediction of Signal Peptide Cleavage Sites with Subsite-Coupled and Template Matching Fusion Algorithm.” *Molecular Informatics* 33: 230–239.

Zheng, Zhi, Youying Chen, Liping Chen, Gongde Guo, Yongxian Fan, and Xiangzeng Kong. 2012. “Signal-BNF: a Bayesian Network Fusing Approach to Predict Signal Peptides.” *Journal of Biomedicine & Biotechnology* 2012: 492174.

Zimmerman, J. M., N. Eliezer, and R. Simha. 1968. “The Characterization of Amino Acid Sequences in Proteins by Statistical Methods.” *J Theor Biol* 21 (2): 170–201.