SignalHsmm - a novel semi-Markov model of eukaryotic signal peptides

Michał Burdukiewicz[1], **Piotr Sobczyk**[2], **Paweł Błażej**[1], **and Paweł Mackiewicz**[1]

[1]**University of Wrocław, Department of Genomics, Poland**
[2]**Wrocław University of Technology, Department of Mathematics, Poland**

ABSTRACT

signalHsmm using universal, transparent and extendable model of signal peptides, predicts their presence in eukaryotic proteins. Signal peptides play important role in cell biology as they are responsible for targeting of proteins to endomembrane system and their export outside the cell. Proteins equipped with signal peptides play crucial roles in metabolism, maintenance of tissue structure, immune response and regulation of other organismal functions. They may be recognized with adequate accuracy using learning systems, that are usually 'black-box' models. Without an ability to trace decision rules, researchers cannot discover which parameters are responsible for predictions. Therefore, we designed a new, more universal probabilistic model for eukaryotic signal peptides, which includes knowledge about their organization, amino acid composition and variability. The proposed approach is based on hidden semi-Markov models (HSMMs) and uses intrinsic knowledge about signal peptides. The big advantage of the algorithm is its extensibility. By using the n-grams (k-mers), we showed that the general model can yield not only better results than other software but also more information about features of signal peptides. Our model showed the largest AUC = 0.97 in comparison to other signal peptide predictor even though it was trained on very small data sets. Moreover, it proved to be very stable regardless of types of learning data sets. Therefore, our model does not need to be permanently retrained with the continuous expansion of sequence databases. It should be emphasized, that our model is able to recognize signal peptides from medically significant malaria parasites *Plasmodium* and their relatives more accurately (with AUC = 0.92) than popular programs (0.84). The web-server of signalHsmm is available at http://smorfland.uni.wroc.pl/signalhsmm.

Keywords: eukaryote; hidden semi-Markov models; n-gram; signal peptide prediction

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 (β galactosidase, pepsins) (Hofmann and Schultz, 1991), maintenance of tissue structure (collagen) (Chan et al., 2001), immune response (interferons, interleukins) (Zhang et al., 2005) and regulation of other organismal functions (prolactin, glucagon) (Huang et al., 2010). Moreover, passing proteins through the endomembrane system is important for their correct folding and posttranslational modification such as glycosylation and phosphorylation.

Despite the low sequence homology between signal peptides (Ladunga, 1999), some general architecture were proposed (Izard and Kendall, 1994; Voss et al., 2013) - Fig. 1. 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 the polypeptide during its translocation through membrane based on the positive-inside rule (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 α-helix. The third part of a signal peptide is a polar, but uncharged c-region. It is usually 6 residues long and ends with a cleavage site, in which the signal peptidase cleaves the signal peptide, during or after translocation of the protein into the 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).

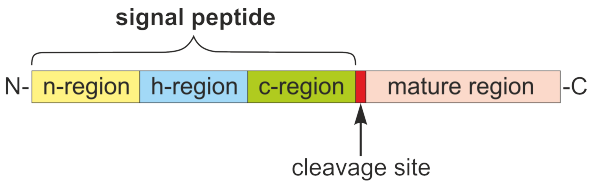

Figure 1. The organization of typical signal peptide. The figure is not drawn in scale.

Some data indicate that signal peptides may be universal. It was found, for example, that even bacterial signal peptides targeted correctly transgenic proteins to plant (Moeller et al., 2009) or mammalian secretory systems (Nagano and Masuda, 2014). On the other hand, signal peptides show great variation and the description presented above (Fig. 1) refers to the most 'typical' signal peptides. There are exceptionally long signal peptides, which fulfill more sophisticated roles (Hiss and Schneider, 2009). 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). The signal peptide from midkine contributing to the tumor progression contains epitopes recognized by CD4+ T cells (Kerzerho et al., 2013). The functional significance of these targeting signals makes that the prediction of signal peptide-containing proteins is also an important step in the drug development (Zhang et al., 2005; Neto Ade et al., 2012; Moeller et al., 2010).

Software predicting signal peptides

Although many experimental methods determining the subcellular localization of proteins were devised, they are time consuming and laborious. Therefore, signal peptides became the subject of many computational programs to their prediction. Many software incorporates 'black-box' models, such as: neural networks (Petersen et al., 2011), support vector machines (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 (Zhang et al., 2014; Hiller et al., 2004). Others (Phobius, Philius and SignalP 3.0) use hidden Markov models (HMMs) (Käll et al., 2004; Reynolds et al., 2008; Bendtsen et al., 2004), which try to reflect structure of signal peptides regions in their limited probabilistic frameworks. The HMMs they use, 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 (Nielsen and Krogh, 1998) and found that the the length distribution for every region was not geometric (Fig. 2). Moreover, the commonly used rigid scheme of signal peptide's organization (Fig. 1) does not describe extremely long or short peptides. Theoretically, HMMs that describe the atypical signal peptides could be developed to consider also the unusual structures, but such probabilistic frameworks have not yet been implemented.

All programs used in signal peptide recognition are trained on real protein sequences. Therefore, they succeed in the recognition of peptides similar to those in the learning set but fail in the case of artificial signal peptides. Such peptides are designed to increase effectiveness of protein secretion (Futatsumori-Sugai and Tsumoto, 2010). They are especially important in industrial applications to increase yield of proteins. Therefore, only explicit knowledge about the organization of signal peptides allows creating sequences that will be the most efficient in the export of proteins (Ng and Sarkar, 2013). Signal peptides have also an important application in gene therapy. Mimicking the natural mechanism of protein export, artificial signal peptides with tumor epitopes increase the antitumor

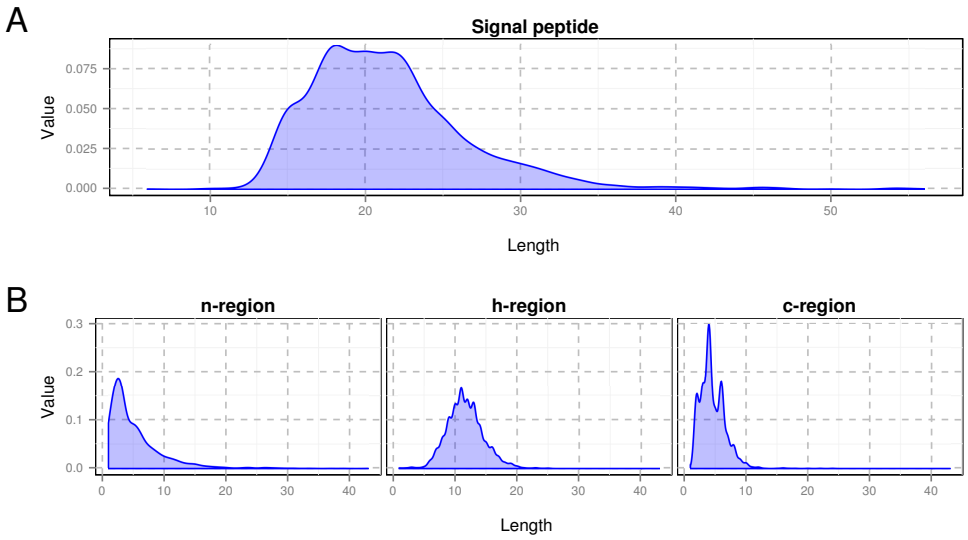

Figure 2. Distribution of lengths of signal peptides (A) and their regions (B) expressed in the number of amino acid residues. The data was extracted from 2 589 signal peptide sequences derived from UniProt database (see **Data selection** in **Methods**).

immune response (He et al., 2003). Such epitopes must be properly inserted into a signal peptide without decreasing its secretion properties through disruption of the regional structure. Instead of time-consuming and expensive laboratory experiments, it would be very useful to survey in silico many artificial peptides to select the ones that would fulfill the designed role.

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 peptides' models because their regions are in fact characterized by specific features of amino acid residues and not by the simple occurrence of specific amino acids. In addition to this, such sparse encoding enforces larger data sets, which hinders their management and analysis (Lin et al., 2002). Therefore, we elaborated a new approach based on hidden semi-Markov models using grouping of amino acids into physicochemical groups characteristic of signal peptides. The new methods proved better in comparison to the current software.

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 characteristics. The pre-processed sequences were further analyzed by the heuristic algorithm, which determines borders between three characteristic signal peptide regions Nielsen and Krogh (1998). We refined some region recognition criteria 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 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

Eukaryotic protein sequences and their annotations were properly prepared according to the literature of the subject and downloaded from UniProt database release 2015_06. The positive set contained 2 589 sequences with an experimentally confirmed signal peptide, containing its start and cleavage site information. Sequences with more than one cleavage site were excluded from the final data

set. The negative set comprised 152 272 sequences without any signal peptide annotation. Protein sequences with ambiguous symbols: X, J, Z, B and selenocysteine (U) were removed from the final sets.

Clustering of amino acids

SignalHsmm is not the first software which uses amino acid encoding in signal peptide prediction. BLOMAP (Maetschke et al., 2005) also employed the similar strategy but considered only substitution matrices. However, we applied different approach. We clustered amino acids using four properties relevant for the architecture of signal peptide: their hydrophobicity, frequency in α-helices, polarity and size. The high hydrophobicity is determinant of the h-region, whose α-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 cleavage site (Palzkill et al., 1994).

Table 1. Properties used in amino acid clusterization.

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

We considered in total 13 amino acid scales from AAIndex database (Kawashima et al., 2008) (Tab. 1). We selected one scale per 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 encodings, we performed a 5-fold cross-validation training a new instance of signalHsmm 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 probability higher than 0.5. Very small variance of performance measures (for example see Tab. 4) confirms credibility of cross-validation.

Hidden semi-Markov model

Hidden semi-Markov model (HSMM) is an extension of hidden Markov model (HMM). Let us first briefly describe the idea of the HMM. Suppose we have a sequence of observations, e.g. amino acids, and we are interested in understanding an underlying cause of their occurrence. HMM aims to answer that question assuming a specific and yet flexible structure of the problem. HMM consists of two stochastic processes. The first is a discrete Markov chain $X_{t=1}^{T}$ on the set of so called hidden states $\{S_1, \ldots S_n\}$. They are "the cause" of the observations. At every step t, hidden state might change according to a transition matrix $A = (a)_{i,j=1}^{n}$, where $a_{i,j} = \mathscr{P}(X_{t+1} = S_j | X_t = S_i)$. In our application, hidden states are signal peptide regions. The second process $E_{t=1}^{T}$ is an observation process defined on the set of possible observations $\{O_1, \ldots, O_m\}$,. They are assumed to occur independently conditionally on the hidden state which emits it. Their distribution probabilities are given by a matrix B, $b_{i,k} = \mathscr{P}(E_t = O_k | X_t = S_i)$. In our case, observations are (encoded) amino acids. The main goal in signalHsmm is to find for a given peptide most probable regions boundaries. This is achieved with Viterbi algorithm. For a good reference on HMM see (Rabiner, 1989).

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., 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 et al., 2004). However, it is still not flexible enough because the empirical regional length distributions (see Fig. 2) are difficult to capture in this way.

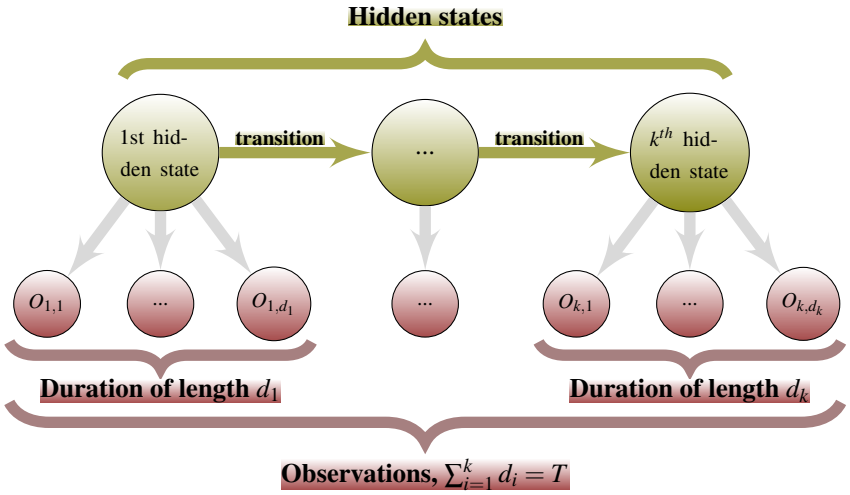

Figure 3. General scheme of hidden semi-Markov model.

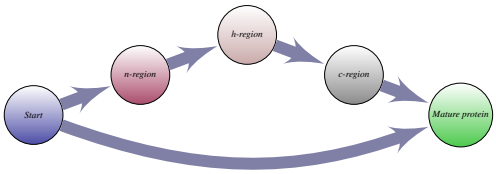

Figure 4. The diagram of signalHsmm.

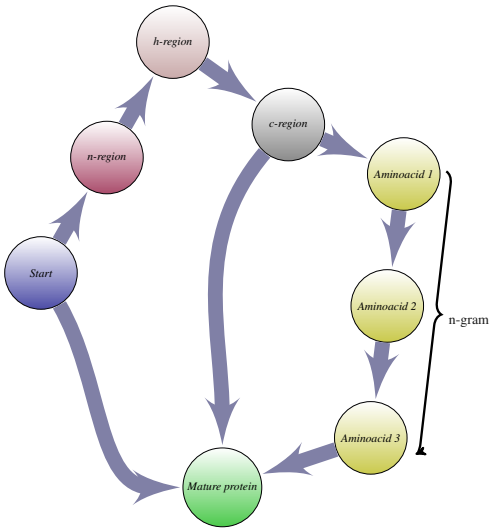

Figure 5. The diagram of signalHsmm extended with the n-gram cleavage site model.

The model we used is Hidden semi-Markov Model(HSMM) (Yu, 2010). It extends HMM by allowing any given hidden state duration distribution (Fig. 3). In addition to matrices A and B, the model is given by probabilities of duration length in hidden states.

$$\mathscr{P}(\text{duration in state} = d | \text{state is } S_i), \;\; i = 1,\ldots,n, \;\; d = 1,\ldots,D$$

where D is the maximum allowed duration. As our datasets are of reasonable size and D is small – around 30, computational effort is not much higher than in the regular HMM.

Our model has a very specific structure. The hidden states represent signal peptide regions. Almost all entries in the transition matrix A are zeros because regions are sequential. Possible transitions are depicted as arrows in Fig. 4. Probabilities of observations for the hidden states and hidden states durations were estimated from training data. The advantage of HSMM model is not only a better performance but also its straightforwardness. Fig. 4 is easy to interpret for a researcher without any mathematical background.

Extension of n-grams

Hidden semi-Markov models may be flexibly enhanced by adding additional hidden states. To improve our model, we added few supplementary states representing specific motives that might occur in the proximity of cleavage site. The structure of cleavage sites, more conserved than other parts of signal peptide (Hiller et al., 2004), may be mirrored by n-grams (k-mers), short vectors of n characters derived from input sequences. Using the biogram software (Burdukiewicz et al., 2015), we extracted n-grams from cleavage sites of signal peptides. The analyzed sequences were already

encoded using amino acid classification providing the best sensitivity of the general model (see Tab.2. Selected n-grams representing less common cleavage site motifs were included in the HSMM model as alternative paths at the end of c-region (Fig. 5).

RESULTS AND DISCUSSION

Cross-validation

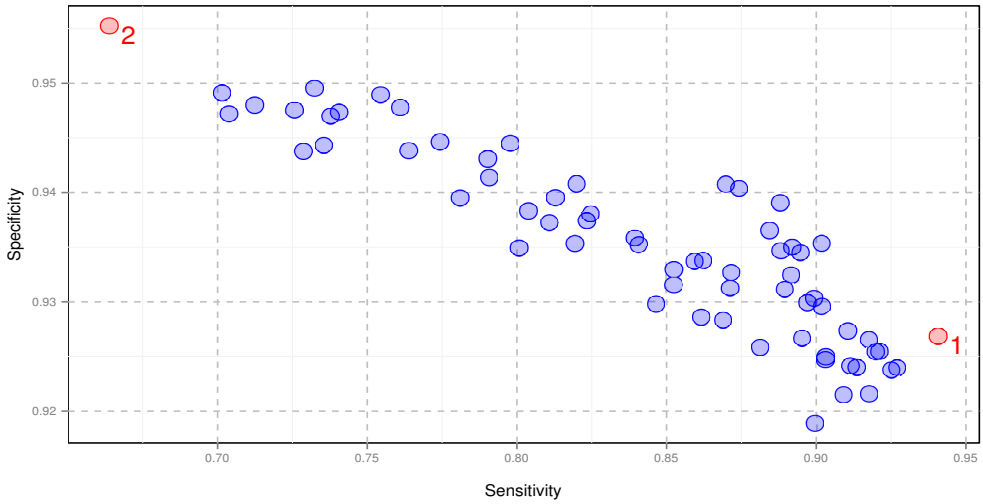

Figure 6. Sensitivity and specificity of amino acid encodings after cross-validation. 1 indicates the encoding providing the best sensitivity (AUC = 0.9683, MCC = 0.8677), whereas 2 means the encoding providing the best specificity (AUC = 0.9338, MCC = 0.6474).

Table 2. The best sensitivity (final) encoding.

Groups
D, E, H, K, N, Q, R
G, P, S, T, Y
F, I, L, M, V, W
A, C

Table 3. The best specificity encoding.

Groups
A, E, K, Q, R
D, G, N, P, S, T
C, H, I, L, M, V
F, W, Y

We used four performance measures to evaluate results of cross-validation: specificity, sensitivity, Matthew's Correlation Coefficient (ϕ coefficient) and Area Under the Curve (AUC). All encodings provided very good AUC (0.93 – 0.97) and specificity (0.92 – 0.96). The classification of amino acids has the biggest impact on sensitivity, which ranges from 0.66 to 0.94. The final signalHsmm algorithm uses the encoding that yields the highest specificity and Matthew's Correlation coefficient as well as the second best AUC (see Fig. 6).

Table 4. Performance measures for the best encoding. 60 repetitions of cross-validation.

Measure	Mean	SD
AUC	0.9683	0.0024
MCC	0.8677	0.0049
Sensitivity	0.9406	0.0008
Specificity	0.9269	0.0050

Comparison of encodings

We examined in more detail encodings with the best sensitivity and the best specificity (Tab. 2, Tab. 2 and Fig. 7). In both cases, the group 1 tends to contain average-sized polar amino acids. In the best sensitivity encoding, all charged amino acids, both acidic and basic (also weakly basic histidine)belong to this group. These amino acids are nearly absent from h-region and provide very

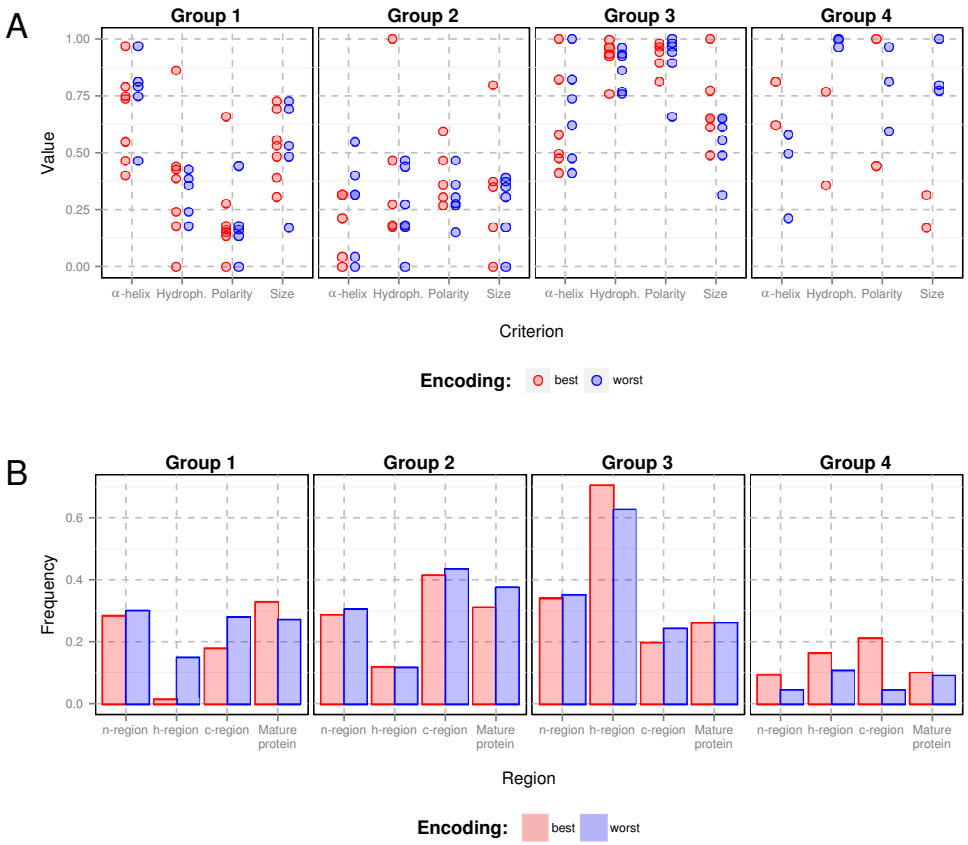

Figure 7. Comparison of amino acid encodings with the best sensitivity and the best specificity in different regions of signal peptide and mature proteins according to: A) the normalized value of property; points represent amino acids; B) frequencies of amino acids in the given region.

good distinction between regions of signal peptide. 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 have a quite low probability of occurrence in α-helix and are very diverse. This group includes polar but uncharged serine and threonine as well as hydrophobic tyrosine, aliphatic glycine and proline. In the best specificity encoding, the polar character of this group is emphasized by the addition of asparagine and aspartic acid. Despite these differences, the distribution of group 2 seems to be comparable in two encodings. The both groupings have strongly non-polar and aliphatic group 3 containing isoleucine, leucine, methionine and valine. The hydrophobic property of this group in the best sensitivity encoding are even more pronounced by the presence of tryptophan and phenylalanine. On the contrary, more polar histidine belongs to the group 3 in the best specificity encoding. Because of the hydrophobic character, this group dominates in the h-region in the case of both amino acid classifications. The fourth group is the most diverse in both encodings. In the case of the best specificity encoding, this group comprises aromatic amino acids: phenylalanine, tryptophan and tyrosine. In contrast, the group 4 in the best sensitivity encoding contains only alanine and cysteine, which both are rather small amino acids and tend to appear in α-helices. This very unique composition seems to be typical of the c-region of signal peptide.

The encoding of amino acid plays crucial role in the recognition of signal peptide but does not affect identification of proteins without signal peptides (compare change in specificity and sensitivity on Fig. 6). The mature protein state in the HSMM model tends to have a more uniform distribution of residues than signal peptide region, which makes it more resistant to changes in amino acid groupings.

Benchmark tests

To provide the honest comparison, we trained our model on 2311 signal peptide-containing sequences deposited in Uniprot until 2010 year (an iteration of signalHsmm called signalHsmm-2010). The data set should correspond to the set used to train SignalP 4.1. Interestingly, our algorithm performed

very well even if it was trained on a very small sets including only 336 sequences collected till 1987 year, just after the first method predicting signal peptide was published (von Heijne, 1986). The signalHsmm-1987 was able to very accurately predict a signal peptide, even better than predictors trained on richer data sets. It indicates that signalHsmm is very stable and is able to recover the structure of signal peptides from even very small data sets (see Tab. 5).

Table 5. Comparison of Area Under the Curve, Sensitivity, Specificity and Matthews Correlation Coefficient for different classifiers.

Software name	AUC	Sensitivity	Specificity	MCC
signalP 4.1 (no tm) (Petersen et al., 2011)	0.9416	**0.9720**	0.9112	0.8848
signalP 4.1 (tm) (Petersen et al., 2011)	0.9673	0.9579	**0.9766**	**0.9347**
PrediSi (Hiller et al., 2004)	0.8949	0.9065	0.8832	0.7899
Phobius (Käll et al., 2004)	0.9509	0.9673	0.9346	0.9024
Philius (Reynolds et al., 2008)	0.9369	0.9533	0.9206	0.8743
signalHsmm-2010	0.9526	0.9533	0.8832	0.8385
signalHsmm-1989	0.9562	0.9626	0.8972	0.8617
signalHsmm-2010 with k-mers	**0.9679**	0.9673	0.9112	0.8799

Table 6. Comparison of Area Under the Curve, H-measure and Matthews Correlation Coefficient for different classifiers considering only proteins belonging to *Plasmodiidae*.

Software name	AUC	Sensitivity	Specificity	MCC
signalP 4.1 (no tm) (Petersen et al., 2011)	0.7928	0.6471	0.9385	0.6185
PrediSi (Hiller et al., 2004)	0.6597	0.3725	**0.9469**	0.4028
Phobius (Käll et al., 2004)	0.7963	0.6765	0.9162	0.5991
Philius (Reynolds et al., 2008)	0.7753	0.6176	0.9330	0.5841
signalHsmm-2010	**0.9340**	0.7941	0.8939	0.6526
signalHsmm-1989	0.9326	0.9216	0.8966	**0.7531**
signalHsmm-2010 with k-mers	0.9334	**0.9902**	0.7989	0.6767

To check universality of our probabilistic model, we also validated it on signal peptides belonging to specific taxonomic groups. As an example, we chose the *Plasmodiidae* family including malaria parasites because of their medical significance. The testing data set extracted from UniProt data base contained 102 sequences with a signal peptide and 358 sequences without it. Interestingly, signalHsmm showed much better performance than other algorithms (Tab. 6). Since our algorithm considers more general decision rules, it is able to recognize also atypical signal peptides.

Conclusions

The architecture of existing signal peptide predicting software is usually opaque which makes analysis of how its decisions are made impossible. Researchers cannot discover which parameters are responsible for the predictions. Considering the biological context of the problem, we see a need for a transparent model of signal peptide. SignalHsmm is the first step in this direction. It provides users good performance as well as interpretability and expendability of predictive model. Aside from the confirmation of the accepted signal peptide model (hydrophibicity of the h-region and polarity of the n-region), we also found that the alanine is one of the most typical amino acids in the c-region in comparison to others. The performance of signalHsmm confirms that properties of signal peptides do not depend on their exact sequence but on the physicochemical features of the amino acids. The flexibility and efficient information recovery makes our model unique among similar software. signalHsmm is also adjustable enough to model properly very specific signal peptides belonging to taxonomic groups which are poorly represented in databases. Moreover, our method can effectively extract information from very small data sets, which in future may lead to new predictors able to recognize atypical signaling sequences.

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.

REFERENCES

Argos, P., Rao, J. K., and Hargrave, P. A. (1982). Structural prediction of membrane-bound proteins. *European Journal of Biochemistry*, 128(2-3):565–75.

Bendtsen, J. D., Nielsen, H., von Heijne, G., and Brunak, S. (2004). Improved prediction of signal peptides: Signalp 3.0. *J Mol Biol*, 340(4):783–95.

Burdukiewicz, M., Sobczyk, P., and Lauber, C. (2015). *biogram: analysis of biological sequences using n-grams*. R package version 1.2.

Chan, D., Ho, M. S. P., and Cheah, K. S. E. (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 Fasman, G. D. (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., Bunster, M., Canales, M., and Gazitua, F. (1992). Hydrophobicity and structural classes in proteins. *Protein Eng*, 5(5):373–5.

Dawson, D. M. (1972). *Size*. Academic Press, New York.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). Biological sequence analysis: probabilistic models of proteins and nucleic acids. cambridge univ.

Eisenberg, D. (1984). Three-dimensional structure of membrane and surface proteins. *Annu Rev Biochem*, 53:595–623.

Fasman, G. D. (1976). *Proteins*, volume 1. CRC Press, Cleveland, 3rd edition.

Futatsumori-Sugai, M. and Tsumoto, K. (2010). Signal peptide design for improving recombinant protein secretion in the baculovirus expression vector system. *Biochem Biophys Res Commun*, 391(1):931–5.

Goldsack, D. E. and Chalifoux, R. C. (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.

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science*, 185(4154):862–4.

He, X., Tsang, T. C., Luo, P., Zhang, T., and Harris, D. T. (2003). Enhanced tumor immunogenicity through coupling cytokine expression with antigen presentation. *Cancer Gene Ther*, 10(9):669–77.

Hegde, R. S. and Bernstein, H. D. (2006). The surprising complexity of signal sequences. *Trends in biochemical sciences*, 31:563–571.

Hiller, K., Grote, A., Scheer, M., Münch, R., and Jahn, D. (2004). Predisi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Research*, 32:W375–W379.

Hiss, J. A. and Schneider, G. (2009). Architecture, function and prediction of long signal peptides. *Brief Bioinform*, 10(5):569–78.

Hofmann, K. J. and Schultz, L. D. (1991). Mutations of the α-galactosidase signal peptide which greatly enhance secretion of heterologous proteins by yeast. *Gene*, 101(1):105–111.

Huang, Y., Wilkinson, G. F., and Willars, G. B. (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, J. W. and Kendall, D. A. (1994). Signal peptides: exquisitely designed transport promoters. *Molecular Microbiology*, 13(5):765–773.

Kapp, K.; Schrempf S.; Lemberg, M. K. D. B. (2000). Post-targeting functions of signal peptides.

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2008). Aaindex: amino acid index database, progress report 2008. *Nucleic Acids Res*, 36(Database issue):D202–5.

Kerzerho, J., Schneider, A., Favry, E., Castelli, F. A., and Maillere, B. (2013). The signal peptide of the tumor-shared antigen midkine hosts cd4+ t cell epitopes. *J Biol Chem*, 288(19):13370–7.

Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, 157(1):105–32.

Käll, L., Krogh, A., and Sonnhammer, E. L. L. (2004). A combined transmembrane topology and signal peptide prediction method. *Journal of molecular biology*, 338:1027–1036.

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.

Lin, K., May, A. C., and Taylor, W. R. (2002). Amino acid encoding schemes from protein structure alignments: multi-dimensional vectors to describe residue types. *J Theor Biol*, 216(3):361–65.

Maetschke, S., Towsey, M., and Bodén, M. (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*, pages 141–150. College Press.

Moeller, L., Gan, Q., and Wang, K. (2009). A bacterial signal peptide is functional in plants and directs proteins to the secretory pathway. *J Exp Bot*, 60(12):3337–52.

Moeller, L., Taylor-Vokes, R., Fox, S., Gan, Q., Johnson, L., and Wang, K. (2010). Wet-milling transgenic maize seed for fraction enrichment of recombinant subunit vaccine. *Biotechnol Prog*, 26(2):458–65.

Nagano, R. and Masuda, K. (2014). Establishment of a signal peptide with cross-species compatibility for functional antibody expression in both escherichia coli and chinese hamster ovary cells. *Biochem Biophys Res Commun*, 447(4):655–9.

Neto Ade, M., Alvarenga, D. A., Rezende, A. M., Resende, S. S., Ribeiro Rde, S., Fontes, C. J., Carvalho, L. H., and de Brito, C. F. (2012). Improving n-terminal protein annotation of plasmodium species based on signal peptide prediction of orthologous proteins. *Malar J*, 11:375.

Ng, D. T. and Sarkar, C. A. (2013). Engineering signal peptides for enhanced protein secretion from lactococcus lactis. *Appl Environ Microbiol*, 79(1):347–56.

Nielsen, H. and Krogh, A. (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.

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

Palzkill, T., Le, Q. Q., Wong, A., and Botstein, D. (1994). Selection of functional signal peptide cleavage sites from a library of random sequences. *Journal of Bacteriology*, 176(3):563–568.

Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). Signalp 4.0: discriminating signal peptides from transmembrane regions. *Nature methods*, 8:785–786.

Ponnuswamy, P. K., Prabhakaran, M., and Manavalan, P. (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, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. pages 257–286.

Radzicka, A., Pedersen, L., and Wolfenden, R. (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.

Rapoport, T. A. (2007). Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. *Nature*, 450(7170):663–9.

Reynolds, S. M., Kall, L., Riffle, M. E., Bilmes, J. A., and Noble, W. S. (2008). Transmembrane topology and signal peptide prediction using dynamic bayesian networks. *PLoS Comput Biol*, 4(11):e1000213.

Shen, H.-B. and Chou, K.-C. (2007). Signal-3l: A 3-layer approach for predicting signal peptides. *Biochemical and biophysical research communications*, 363:297–303.

Szczesna-Skorupa, E., Browne, N., Mead, D., and Kemper, B. (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.

von Heijne, G. (1986). A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res*, 14(11):4683–90.

von Heijne, G. and Gavel, Y. (1988). Topogenic signals in integral membrane proteins. *European Journal of Biochemistry*, 174(4):671–8.

Voss, M., Schröder, B., and Fluhrer, R. (2013). Mechanism, specificity, and physiology of signal peptide peptidase (spp) and spp-like proteases. *Biochimica et biophysica acta*, 1828:2828–2839.

Yu, S.-Z. (2010). Hidden semi-markov models. *Artificial Intelligence*, 174(2):215 – 243. Special Review Issue.

Zhang, L., Leng, Q., and Mixson, A. J. (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, S.-W., Zhang, T.-H., Zhang, J.-N., and Huang, Y. (2014). Prediction of signal peptide cleavage sites with subsite-coupled and template matching fusion algorithm. *Molecular Informatics*, 33:230–239.

Zheng, Z., Chen, Y., Chen, L., Guo, G., Fan, Y., and Kong, X. (2012). Signal-bnf: a bayesian network fusing approach to predict signal peptides. *Journal of biomedicine & biotechnology*, 2012:492174.

Zimmerman, J. M., Eliezer, N., and Simha, R. (1968). The characterization of amino acid sequences in proteins by statistical methods. *J Theor Biol*, 21(2):170–201.