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

The proper localization of proteins in a cell is essential to maintain their desired function. Information about the protein destination is included within the very protein in the form of short peptides called targeting signals. Ones of them are signal peptides, diverse N-terminal sequences, which 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. Moreover, the transport of proteins through the endomembrane system is important for their correct folding and posttranslational modifications.

A common model of classical signal peptides assumes that they start with a positively charged n-region, followed by a hydrophobic h-region and a c-region ended with a cleavage site recognised by a signal peptidase. However, our studies of many protein sequences representing the wide range of diversified taxonomic organisms indicate a variability of signal peptides. Therefore, we designed a new, more univeral probabilistic model for eukaryotic signal peptides, which includes knowledge about their organisation, amino acid composition and variation.

The proposed model is based on hidden semi-Markov models (HSMMs) and use intrinsic knowledge about signal peptides. The big advantage of the algorithm is its extensibility. Using the n-grams (k-mers) we point how the general model can be attuned to yield not only better results, but also more information about signal peptides.

Our model was validated a signal peptide pedictor. It has showed the largest AUC=0.98 in comparison to other software and appeared very stable in the recovery of signal peptides after training even on very small data sets. Thanks to that, our model does not need to be permanently retrained with the continuous expansion of sequence databases. It should be emphasised that our model describes signal peptides from medically significant malaria parasites Plasmodium and their relatives (AUC = 0.92) more accurately than popular programs (0.84).

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

INTRODUCTION

Signal peptides

Proteins of eukaryotes are encoded in nuclear genomes and are synthetized in ribosomes located in the cytosol or bounded by the endoplasmic reticulum. After the translation process, proteins have to be targeted to specific subcellular compartments or exported outside the cell to the extracellular environment. The proper localization of proteins is essential to perform their desired function. Information about protein destination is included within the very protein in the form of short peptides or stretches of amino acid residues called targeting or sorting signals. Ones 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, Golgi apparatus and endosomes. Such proteins can stay inside of these compartments, or can be inserted into cellular membranes or exported outside the cell. Proteins equipped with signal peptides constitute a substantial fraction of the whole proteome. They play crucial roles 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.

Although signal peptides are quite variable, 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 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 has a tendency to form α-helix. The third part of a signal peptide, usually 6 residues long, is a polar, but uncharged c-region ended with a cleavage site recognized by the signal peptidase. 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).

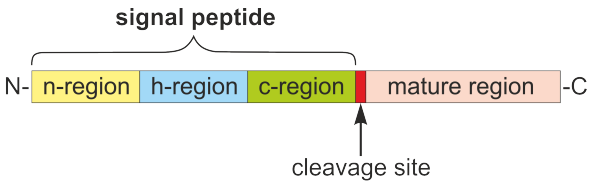

Figure 1. The organization of signal peptide.

During or after translocation of the protein into the lumen of endoplasmic reticulum, the typical signal peptide is cleaved by a signal peptidase (Paetzel et al., 2002) and next degraded by specific proteases, whereas the rest (mature) part of protein stays in the lumen or is passed to other compartments. The cleavage site is characterized by a very variable amino acid composition. It typically contains small and neutral residues at -3 and -1 positions (Palzkill et al., 1994). The site is, however, absent from some membrane proteins in which the first transmembrane-domain acts both as signal peptides and signal anchor (Szczesna-Skorupa et al., 1988).

Some data indicate that signal peptides may be universal. It was found, for example, that even bacterial signal peptides targeted correctly transgenic proteins to the plant (Moeller et al., 2009) or mammalian secretory system (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 also exceptionally long signal peptides, which fulfil more sophisticated roles (Hiss and Schneider, 2009). For example, the fragment of signal peptide from preprolactin probably takes part in the regulation of prolactin secretion. Other examples are signal peptides of MHC class I, which inhibit activity of NK cells. Interesting functions have signal peptides of viral origin, which are involved in the immune evasion or viral life cycle (Kapp, 2000). Such diverse functions are restricted not only to the long signal peptides. The peptides from midkine, a protein contributing to the tumor progression, contains epitopes recognized by CD4+ T cells (Kerzerho et al., 2013). It indicates that the signal peptide may participate in a tumor immunity. 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).

Signal peptide predicting software

Although many experimental methods determining the subcellular localization of proteins were devised, they are time consuming and laborious. Therefore, the development of new approaches in the field of computational biology and bioinformatics is desirable. They are not only a good support, complement or alternative for the experimental methods but also enable to understand rules encoding information about protein targeting, which is contained in the predicted targeting signals. However, many of the present predictors disregard full biological information carrying by signal peptides.

Having functional importance and specific features, signal peptides became the subject of many programs to their prediction based on different methods. The state of the art software for predicting the presence of signal peptides often incorporates 'black-box' models, 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). Such models neither utilize biological information about signal peptides nor work properly on atypical signal peptides.

The other type of software incorporates into the prediction process theoretical knowledge about the structure of signal peptides. Although these programs do not share innate flaws of 'black-box' models, they also demand an improvement. Some of them are based on position matrices or their more abstract variants (Zhang et al., 2014; Hiller et al., 2004). Others (Phobius, Philius and not supported SignalP 3.0) use hidden Markov models (HMMs) (Käll et al., 2004; Reynolds et al., 2008;

Bendtsen et al., 2004), which reflect regions of signal peptides without realizing limitations of this probabilistic framework. However, the HMMs imply geometric distribution for duration of region length. We replicated the rules for extracting regions' boundaries from the first work utilizing HMMs in prediction of signal peptides (Nielsen and Krogh, 1998) and found that the mentioned assumption is not in concordance with the reality because every region revealed the length distribution other than geometric (Fig. 2). Moreover, the commonly used rigid region scheme (Fig. 1) does not describe extremely long or short signal peptides. Theoretically, HMMs that describe the atypical signal peptides could be developed to consider also unusual structures, but such probabilistic frameworks have not still been implemented.

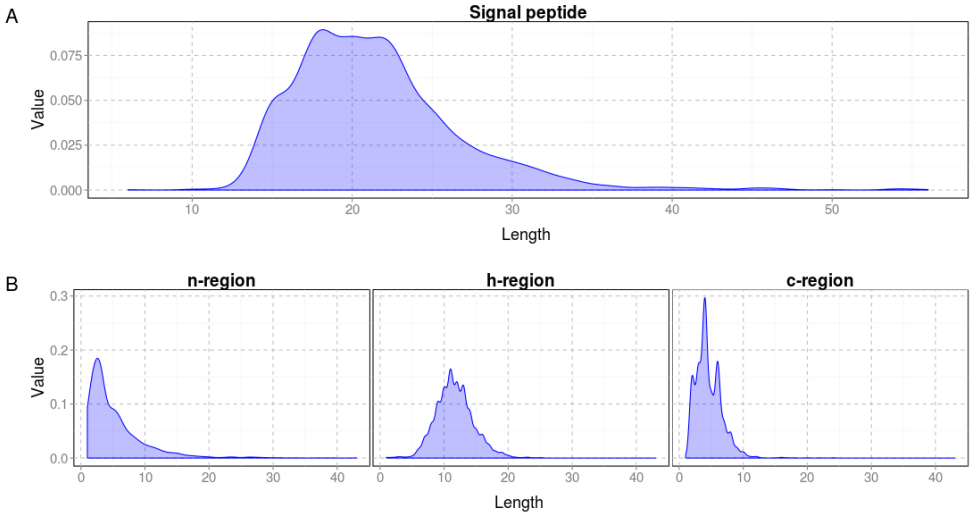

Figure 2. A) Distribution of lengths of signal peptides. B) Distribution of the lengths of signal peptide regions. The measure of the length is a number of amino acids in the subsequence. The data was extracted from 2 589 signal peptide sequences derived from UniProt database (see **Data selection** in **Methods**).

Another feature, shared by majority of the signal peptide predicting software, is the orthogonal encoding of amino acids, in which a vector of 20 digits represents every amino acid. This method of encoding, however, does not preserve relationships between amino acids. The distance between residues is the same regardless of the similarity of their physicochemical properties. Such behaviour is especially undesired in signal peptides models because their regions are defined 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).

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 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 fulfil the designed role.

METHODS

Overview

The functionality of a signal peptide depends not on exact sequence of specific amino acids, but on the physicochemical properties of residues in a given region. Henceforth, the usage of raw amino acid sequences is superfluous and introduces unnecessary information. To utilize this property of

signal peptide recognition, we cluster amino acids into several groups based on the physicochemical features of residues.

The pre-processed sequences are further analyzed by the heuristic algorithm, which determines borders between three characteristic signal peptide regions, the enhanced version of algorithm presented in Nielsen and Krogh (1998). Using the current information from experimentalists, we refined the region recognition criteria.

Next, two models are trained to recognize proteins with and without a signal peptide. The first one is a hidden semi-Markov model, in which each of three signal peptide regions is represented by a different hidden state. The additional fourth hidden state represents mature protein. Each state is described by the frequencies of amino acid groups within that state. The distribution of hidden states durations, the number of amino acids related to each hidden state in signal peptide, is based on the empirical density of region lengths from the training set.

The second model is 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 and its cleavage site. 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 and B were removed from the final sets. Proteins with selenocysteine (U) were also excluded from data set, because there are no records of signal peptides containing this amino acid.

Clustering of amino acids

It is worth to note, that signalHsmm is not a first software to use amino acid encoding in signal peptide prediction. BLOMAP (Maetschke et al., 2005) also employed similar strategy, but considered only encodings based on the substitution matrices. Substitution matrices are computed using protein sequences and knowning how atypical signal peptides are, we had chosen different approach.

Instead amino acids were clustered using several criteria relevant for the architecture of signal peptide: hydrophobicity, frequency in alpha-helices, polarity and size. High hydrophobicity is a determinant of the h-region, the core of signal peptides. Alpha-helix, the secondary structure of the h-region, is probably induced by the positively charged n-region. High polarity as well as smaller size are important features of the residues in the cleavage site (Palzkill et al., 1994).

Table 1. Properties used in clusterization.

Criterion name	Property name
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 alpha-proteins (Cid et al., 1992)
Hydrophobicity	Consensus normalized hydrophobicity scale (Eisenberg, 1984)
Hydrophobicity	Hydropathy index (Kyte and Doolittle, 1982)
Hydrophobicity	Surrounding hydrophobicity in alpha-helix (Ponnuswamy et al., 1980)
Polarity	Polarity (Grantham, 1974)
Polarity	Mean polarity (Radzicka et al., 1988)
Frequency in alpha-helices	Signal sequence helical potential (Argos et al., 1982)
Frequency in alpha-helices	Normalized frequency of N-terminal helix (Chou and Fasman, 1978)
Frequency in alpha-helices	Relative frequency in alpha-helix (Prabhakaran, 1990)

We selected 13 properties from AAIndex database (Kawashima et al., 2008) (see Tab. 1), each attributed to a single criterion. We established 96 permutations of the properties, where each permutation contains only one property associated with a given criterion. Considering all permutations of properties, we created 96 possible clusterings of amino acids using Euclidean distance and Ward's method. Further, we cut the clusterings to create four-group encodings. As expected, some encodings (31%) were identical.

To compare encodings we performed a 5-fold cross-validation training a new instance of signalHsmm on every encoding. We created balanced data sets by subsampling a number of proteins without a signal peptide equal to the number of proteins with signal peptide. The analysis was carried using also redundant encodings, but results are not shown (HSMM is a deterministic algorithm and predictions for identical groupings of amino acids were the same). The cross-validation was repeated 60 times. At first, we planned 160 repetitions to ensure with probability 0.95, that every protein without signal peptide was in the learning set, but very small variance of performance measure allowed us to relax this condition.

Hidden semi-Markov model

Hidden semi-Markov model (HSMM) is an extension of hidden Markov model (HMM). As such, it proved to yield better performance in many fields of bioinformatics including gene recognition (Pachter et al., 2002), protein secondary structure prediction (Aydin et al., 2006) or automatic detection of heartbeats (Pimentel et al., 2014). Let us first briefly describe the idea that lies behind HMM. Suppose you have a sequence of observations e.g. amino acids, and you are interested in understanding underlying cause of their occurrence. HMM aims to answer that question by assuming a specific, yet flexible, structure of the problem. HMM consists of two stochastic processes. First is 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 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. Second process $E_{t=1}^{T}$ is observation process, defined on the set of possible observations $\{O_1, \ldots, O_m\}$. Observations are assumed to occur independently. Their distribution depends only on hidden state which emits it. 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 (degenerated) amino acids. 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 regular HMM, hidden state duration – number of observations emitted by the hidden state, has geometric distribution. (Durbin et al., 1998) shows how to extend it for some different distributions without significant increase in computational complexity. Similar ideas were used for signal peptide recognition, for example in (Käll et al., 2004). It is, however, still not flexible enough. Empirical regional length distributions (see Fig. 2) are difficult to capture in that way.

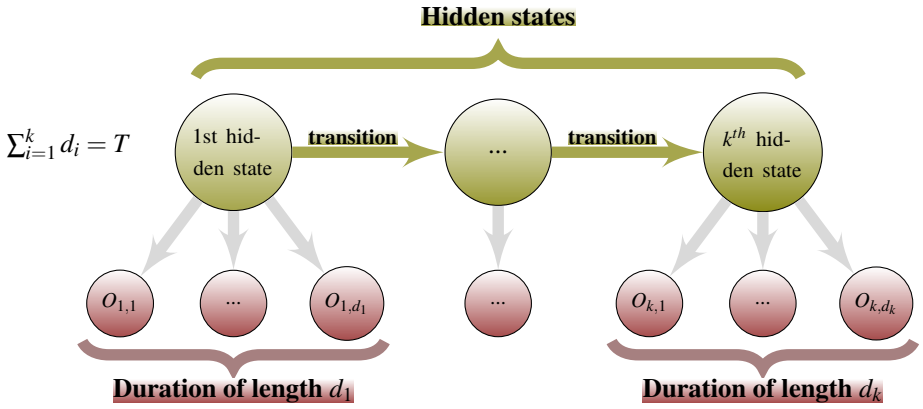

Figure 3. General scheme of hidden semi-Markov model.

The model we use is Hidden semi-Markov Model(HSMM) (Yu, 2010). It extends HMM by allowing any given hidden state duration distribution. Diagram of HSMM is presented in Fig. 3. In addition to matrices A and B, 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 maximum allowed duration. As our datasets are of reasonable size, D is small – around 30, computational effort is not much higher than with the regular HMM.

The model we use has a very specific structure. Hidden states are signal peptide regions. Almost all entries in transition matrix A are zeros, because regions are sequential. Possible transitions are depicted as arrows in Fig. 4. Probabilities of observations for 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.

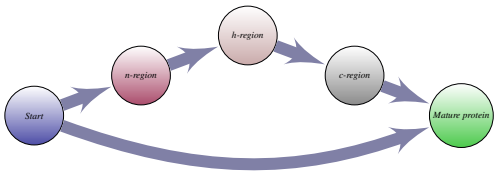

Figure 4. The diagram of signalHsmm.

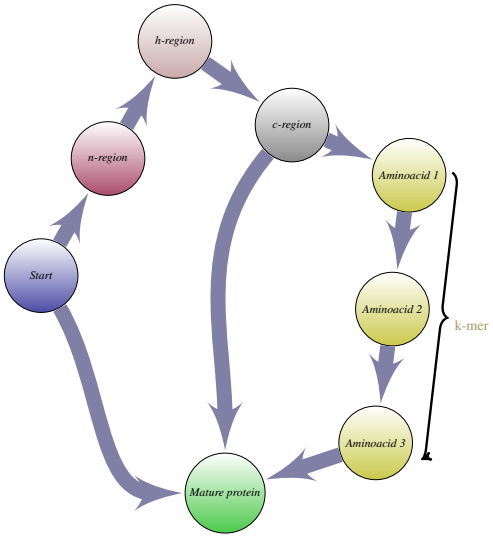

Figure 5. The diagram of signalHsmm extended with the n-gram cleavage site model.

n-gram extension

HSMM can be further extended to use some specific motives that might occur in proximity of cleavage site. Such an extension requires defining additional hidden states. To find such motives we analyze n-grams (k-mers) – vectors of n characters derived from input sequences. The rigid structure of n-grams may mirror cleavage sites, which are more conservative than other parts of signal peptide (Hiller et al., 2004). Scheme of extended HSMM can be seen in Fig. 5.

Using the biogram software (Burdukiewicz et al., 2015) we extracted n-grams from cleavage sites of signal peptides. XXX n-grams were further used in the analysis

RESULTS

Cross-validation

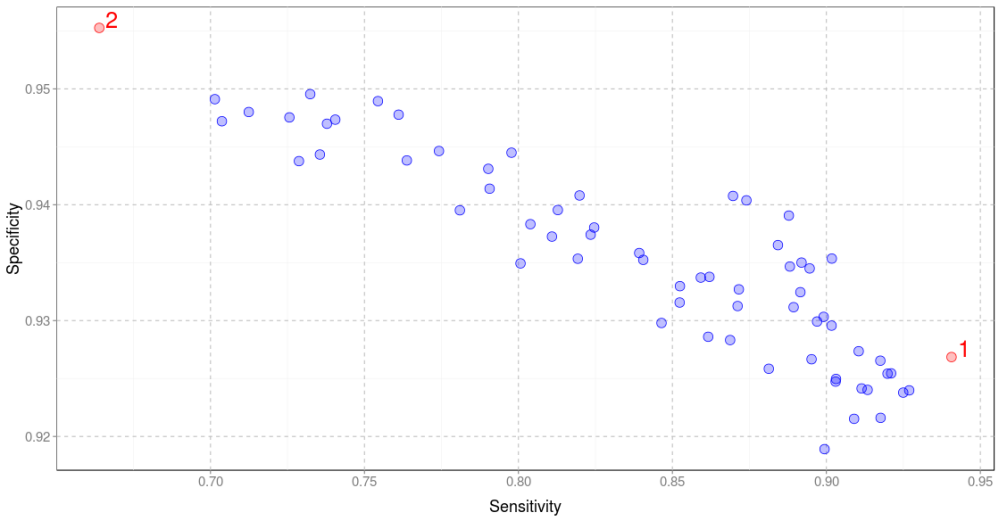

Figure 6. Results of cross-validation. 1. The encoding providing the best sensitivity (AUC = 0.9683, MCC = 0.8677). 2. The encoding providing the best specificity (AUC = 0.9338, MCC = 0.6474).

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 provide very good AUC (0.93 – 0.97) and specificity (0.92 – 0.96). The encodings have 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

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 worst sensitivity encoding.

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

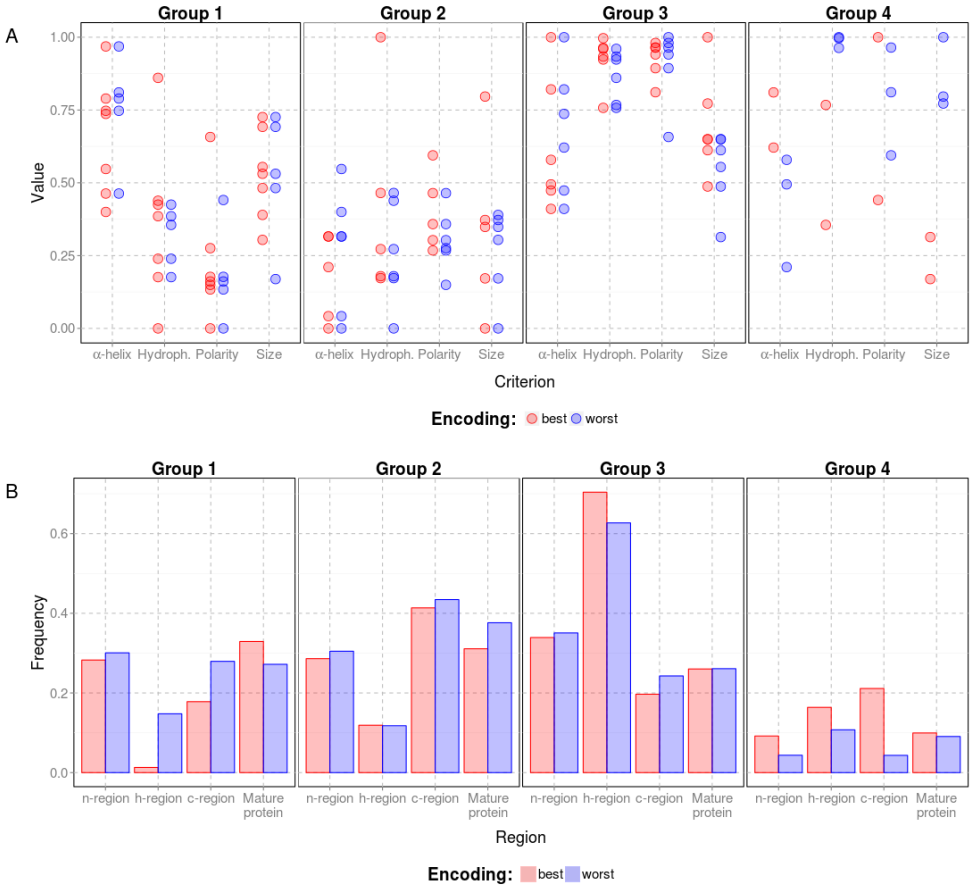

Figure 7. A comparision of encodings. A) Properties of amino acids in the best and the worst sensitivity encoding. B) Frequencies of amino acids beloning to the encodings in regions of signal peptide and mature proteins.

second best AUC (see Fig. 6).

Comparision of encodings

We examined more deeply encodings with the best and the worst sensitivity. In both cases Group 1 tends to contain polar amino acids typical with the average size of residuals. In the best encoding, all charged amino acids, both acids and even weak bases as histidine, also belong to this group. This group, nearly absent in h-region, provides very good distinction between parts of signal peptide for best encoding. In the worst encoding, where polar and charged character of the Group 1 is not that explicit, the difference in frequency between regions is also less visible.

The amino acids belonging to the Group 2 have quite low probability of occuring in an alpha-helix, but are still very diverse, but all . Polar but uncharged serine and threonine neighbour with alyphatic glycine and proline. In the worst encoding the polar character of the group is emphasized by the addition of asparagine. The best grouping contains also highly hydrophobic tyrosine. Despite these differences, the frequency of Group 2 seems to be comparable between two encodings.

The both groupings have strongly non-polar and aliphatic group 3 (see Tab. 2, Tab. 2 and Fig. 7) containing isoleucine, leucine, methionine and valine. Hyrophobic properties of this group in the best encoding are even more pronounced by the presence of tryptophan and phenylalanine. On

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

the contrary, to the Group 3 in the worst encoding belong also more polar histidine. Hence the hydrophobic character of this group, it is dominant over the h-region in case of both amino acid classifications.

The fourth group is the most diverse among both encodings. The key is obvious in a case of the worst encoding. Phenylalanine, tryptophan and tyrosine are aromatic, highly hydrophobic amino acids. In contrast, the Group 4 in the best encoding contains only alanine and cysteine, which both are rather small amino acids and tend to appear in alpha-helices. This very unique composition seems to be typical for the c-region of signal peptide.

Benchmark test

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 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)	0.9424	0.9720	0.9128	0.8859
signalP 4.1 (tm) (Petersen et al., 2011)	0.9675	0.9579	0.9771	0.9353
PrediSi (Hiller et al., 2004)	0.8959	0.9065	0.8853	0.7919
Phobius (Käll et al., 2004)	0.9515	0.9673	0.9358	0.9033
Philius (Reynolds et al., 2008)	0.9376	0.9533	0.9220	0.8755
signalHsmm-2010	0.9429	0.9393	0.8761	0.8166
signalHsmm-1989	0.9455	0.9486	0.8945	0.8440

To check universality of our probabilistic model, we also validated it on signal peptides belonging to specific taxonomic groups. As the example, we chose the Plasmodiidae family including malaria parasites. The testing data set extracted from UniProt data base contained 111 proteins with a signal peptide and 361 without it. We repeated the benchmark scheme for different classifiers (Tab. 6).

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

DISCUSSION

Despite the presence of signal peptide predicting software, there are no exhaustive studies devoted to the analysis of the signal peptide structure and taxonomical variability. Our research focuses on a transparent model of signal peptide instead of a 'black-box' solution.

The good 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. Although this finding may seem to offer a minimal improvement when we consider the most popular signal peptides, this generalization pays off when we analyze more specific cases, as signal peptides of *Plasmodiidae*.

Aside from a confirmation of the accepted signal peptide model (hydrophibicity of the h-region, polarity of the n-region), we also found new relationships. We found the alanine is one of the most typical amino acids for the c-region and occurs there far more often than in the other parts of signal peptides and mature proteins. The presence of alanine in the signal peptide cleavage site von Heijne (1986) cannot explain such overrepresentation of group containing alanine in the c-region.

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

The proposed algorithm recognizing various signal peptides will reduce costs and speed up searching appropriate artificial signal peptides designed for protein secretion in expensive and time-consuming laboratory experiments. Moreover, the algorithm can be applied in recognition any sequence types on the nucleotide and amino acid levels, represented by small sets. Accessible as a stand-alone and browser-based software, its useful both for researches interested in the analysis of signal peptides.

ACKNOWLEDGMENTS

PS and PB devised the HSMM model of signal peptide. MB and PS co-wrote the software. MB gathered data and carried the analysis. MB, PS and PM drafted the manuscript. All authors revised the manuscript and approved the final version.

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.

Aydin, Z., Altunbasak, Y., and Borodovsky, M. (2006). Protein secondary structure prediction for a single-sequence using hidden semi-markov models. *BMC Bioinformatics*, 7(1):178.

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.

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.

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.

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.

Pachter, L., Alexandersson, M., and Cawley, S. (2002). Applications of generalized pair hidden markov models to alignment and gene finding problems. *J. Comput. Biol*, 9:389–399.

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.

Pimentel, M., Santos, M., Springer, D., and Clifford, G. (2014). Hidden semi-markov model-based heartbeat detection using multimodal data and signal quality indices. In *Computing in Cardiology Conference (CinC), 2014*, pages 553–556.

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.