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).

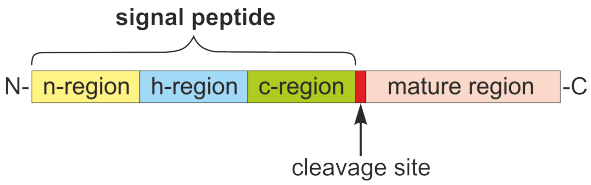

Figure 1. The organization of signal peptide.

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 (not supported SignalP 3.0, Philius and Phobius) use hidden Markov models (HMMs) (Käll et al., 2004; Reynolds et al., 2008),

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.

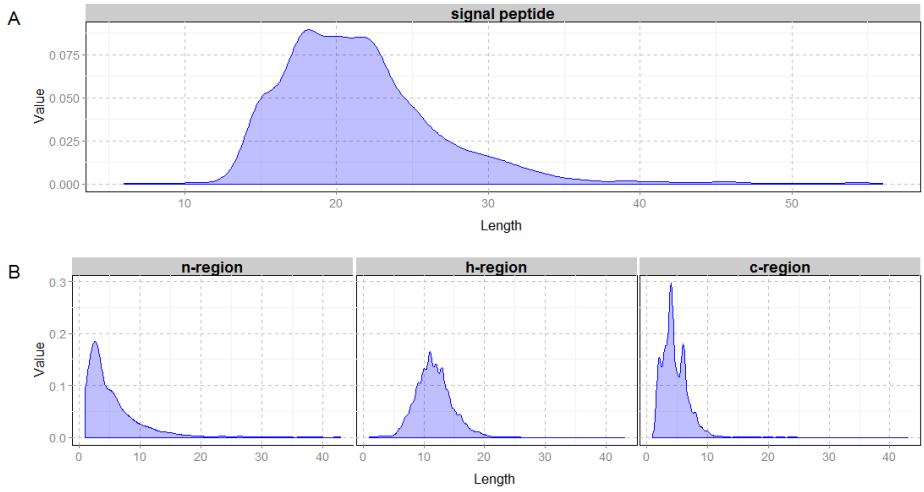

Figure 2. A) Distribution of lengths of signal peptides. B) Distribution of the lengths of signal peptide regions. The data was extracted from 2 589 signal peptide sequences derived from UniProt database (see **Data selection** in **Methods**).

Another feature, shared by majority of the signal peptide predicting software, is 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 properties 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

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

Criterion name	Property name
Size	Size
Size	Molecular weight
Size	Residue volume
Size	Bulkiness
Hydrophobicity	Normalized hydrophobicity scales for alpha-proteins
Hydrophobicity	Consensus normalized hydrophobicity scale
Hydrophobicity	Hydropathy index
Hydrophobicity	Surrounding hydrophobicity in alpha-helix
Polarity	Polarity
Polarity	Mean polarity
Frequency in alpha-helices	Signal sequence helical potential
Frequency in alpha-helices	Normalized frequency of N-terminal helix
Frequency in alpha-helices	Relative frequency in alpha-helix

Table 1. Properties used in clusterization.

We selected 13 properties from AAIndex database (aaindexcitation) (see Table 1), each attributed to a single criterion. Considering all combinations of single properties for every criterion, we created 96 possible clusterings of amino acids using Euclidean distance and Ward's method.

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 cross-validation was repeated XXX times, to ensure that every negative protein was tested at least once with XXX probability.

Hidden semi-Markov model

Tu troche matematyki o ukrytych modelach Markowa (Rabiner, 1989).

RESULTS

You may want to separate results, discussion and conclusion, according to your needs.

Please submit the final pdf file via EasyChair to the GCB'15 program committee by June 30, 2015.

DISCUSSION

ACKNOWLEDGMENTS

Thank you for your support!

REFERENCES

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.

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. Futatsumori-Sugai, Mutsumi Tsumoto, Kouhei Biochem Biophys Res Commun. 2010 Jan 1;391(1):931-5. doi: 10.1016/j.bbrc.2009.11.167. Epub 2009 Dec 5.

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. He, Xianghui Tsang, Tom C Luo, Phoebe Zhang, Tong Harris, David T England Cancer Gene Ther. 2003 Sep;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. Hiss, Jan A Schneider, Gisbert England Brief Bioinform. 2009 Sep;10(5):569-78. doi: 10.1093/bib/bbp030. Epub 2009 Jun 17.

Hofmann, K. J. and Schultz, L. D. (1991). Mutations of the alpha-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.

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. Kerzerho, Jerome Schneider, Aurelie Favry, Emmanuel Castelli, Florence Anne Maillere, Bernard J Biol Chem. 2013 May 10;288(19):13370-7. doi: 10.1074/jbc.M112.427302. Epub 2013 Apr 3.

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.

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. Lin, Kuang May, Alex C W Taylor, William R England J Theor Biol. 2002 Jun 7;216(3):361-65.

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, Lorena Gan, Qinglei Wang, Kan England J Exp Bot. 2009;60(12):3337-52. doi: 10.1093/jxb/erp167. Epub 2009 Jun 2.

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. Moeller, Lorena Taylor-Vokes, Raye Fox, Steve Gan, Qinglei Johnson, Lawrence Wang, Kan Biotechnol Prog. 2010 Mar-Apr;26(2):458-65. doi: 10.1002/btpr.326.

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. Nagano, Ryuma Masuda, Kazuhiro Biochem Biophys Res Commun. 2014 May 16;447(4):655-9. doi: 10.1016/j.bbrc.2014.04.060. Epub 2014 Apr 19.

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. Neto, Armando de Menezes Alvarenga, Denise A Rezende, Antonio M Resende, Sarah S Ribeiro, Ricardo de Souza Fontes, Cor J F Carvalho, Luzia H de Brito, Cristiana F Alves England Malar J. 2012 Nov 15;11:375. doi: 10.1186/1475-2875-11-375.

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. Ng, Daphne T W Sarkar, Casim A Appl Environ Microbiol. 2013 Jan;79(1):347-56. doi: 10.1128/AEM.02667-12. Epub 2012 Nov 2.

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. Paetzel, Mark Karla, Andrew Strynadka, Natalie C J Dalbey, Ross E GM63862/GM/NIGMS NIH HHS/ Chem Rev. 2002 Dec;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.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. pages 257–286.

Rapoport, T. A. (2007). Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. *Nature*, 450(7170):663–9. Rapoport, Tom A England Nature. 2007 Nov 29;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. Reynolds, Sheila M Kall, Lukas Riffle, Michael E Bilmes, Jeff A Noble, William Stafford P41-RR11823/RR/NCRR NIH HHS/ R01-EB007057/EB/NIBIB NIH HHS/ PLoS Comput Biol. 2008 Nov;4(11):e1000213. doi: 10.1371/journal.pcbi.1000213. Epub 2008 Nov 7.

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. 3422456[pmid] Proc Natl Acad Sci U S A.

von Heijne, G. and Gavel, Y. (1988). Topogenic signals in integral membrane proteins. *European Journal of Biochemistry*, 174(4):671–8. von Heijne, G Gavel, Y GERMANY, WEST Eur J Biochem. 1988 Jul 1;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.

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, Lei Leng, Qixin Mixson, A James CA 70394/CA/NCI NIH HHS/ CA 96984/CA/NCI NIH HHS/ England J Gene Med. 2005 Mar;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.