Predicting eukaryotic signal peptides using hidden Markov models

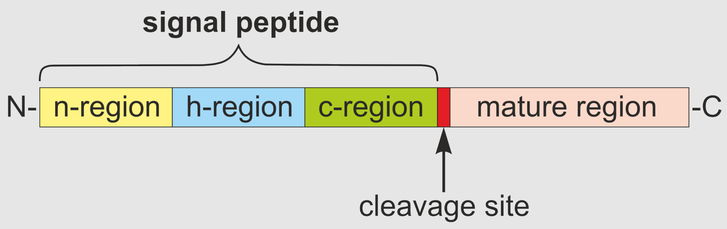
Michał Burdukiewicz[1]*, Piotr Sobczyk[2], Paweł Błażej[1], Paweł Mackiewicz[1]
*michalburdukiewicz@gmail.com

[1]University of Wrocław, Department of Genomics, [2]Wrocław University of Science and Technology, Faculty of Pure and Applied Mathematics

Introduction

The computational methods for the recognition of signal peptides accurately identify typical peptides, well-represented in protein databases (Petersen et al., 2011). However, these algorithms are not general enough to predict more unique signal peptides, for example those present in proteins from parasites belonging to the phylum Apicomplexa.
Apicomplexans are characterized by a strongly AT-biased genomes and resulting from that the specific amino acid composition of coded proteins, which hinders their computational detection by general predictors.
Nevertheless, members of Apicomplexa have a great medically significance, especially *Plasmodium*, a malaria agent.

Signal peptide architecture



- n-region: mostly basic residues (Nielsen and Krogh, 1998),
- h-region: strongly hydrophobic residues (Nielsen and Krogh, 1998),
- c-region: a few polar, uncharged residues (Jain et al., 1994).

Hidden semi-Markov model (HSMM) of a signal peptide

Assumptions of the model:

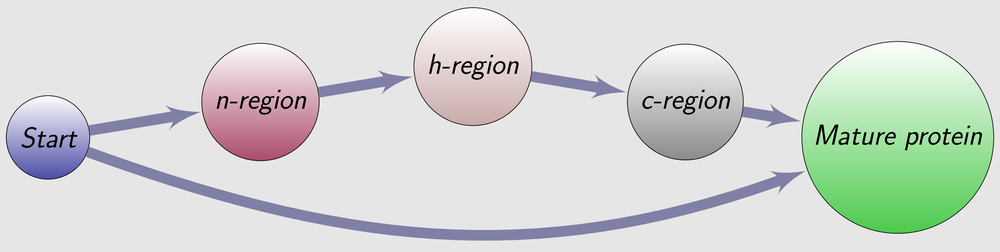
- the observable distribution of amino acids arises due to being in a certain region (state),
- a duration of the state (the length of given region) is modeled by a probability distribution (other than geometric distribution as in typical hidden Markov models).

Training of signalHsmm

1. Removal of atypical (non-standard amino acids, more than one cleavage site) or poorly annotated records from data set of proteins with signal peptide from UniProtKB 2014_07 (after purification data set contains 3816 eukaryotic proteins with experimentally confirmed signal peptides and 9795 without signal peptide).
2. Determination of n-, h-, c-regions by the heuristic algorithm.
3. Reduction of dimensionality by aggregating amino acids to several physicochemical groups.
4. Calculation of the amino acid group frequency in each region and the average length of the region.
5. Training of two HSMM models for proteins with and without signal peptide.

Signal peptide prediction

During the test phase, each protein is fitted to two HSMMs representing respectively proteins with and without signal peptides. The probabilities of both fits and predicted cleavage site constitute the software output.



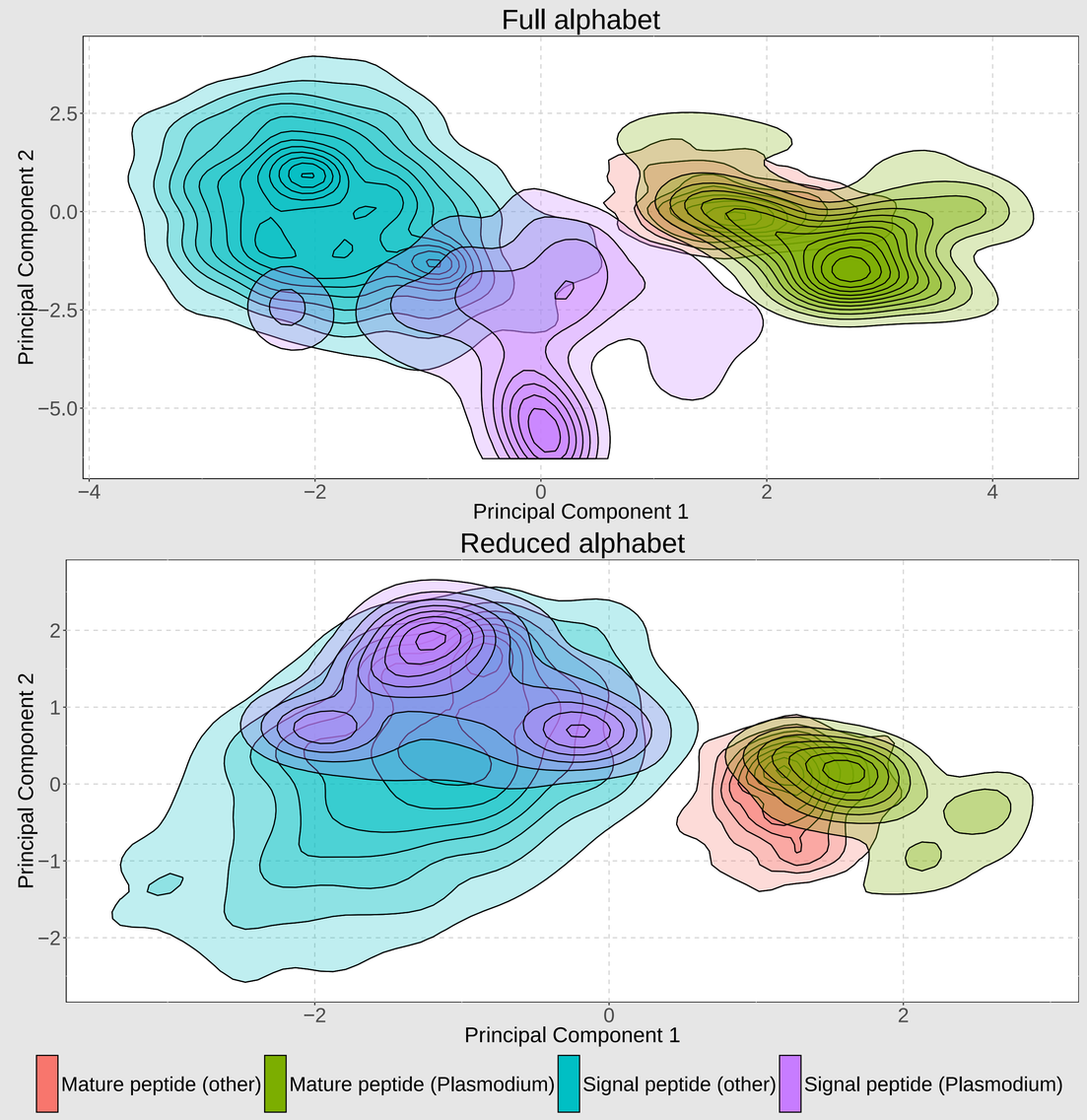
Classification of amino acids used by signalHsmm

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

Availability and funding

signalHsmm web server: `www.smorfland.uni.wroc.pl/signalhsmm`
This research was partially funded by the KNOW Consortium and National Science Center (2015/17/N/NZ2/01845).
Find our poster online: `http://github.com/michbur/GCB2016`

PCA of signal peptides and mature proteins



Countour plot of first two components in Pricincipal Compoment Analysis of amino acid frequency. The signal peptides from malaria and other taxons differ significantly when the full amino acid alphabet is employed. After the reduction of the alphabet, the signal peptides group together despite their origin.

Benchmark with other predictors of signal peptides

Benchmark data set: 51 proteins with signal peptide and 211 proteins without signal peptide from members of *Plasmodiidae*.

signalHsmm1987: trained on data set of 496 eukaryotic proteins with signal peptides added before year 1987.

signalHsmm2010: trained on data set of 3676 eukaryotic proteins with signal peptides added before year 2010.

	Sensitivity	Specificity	MCC	AUC
signalP 4.1 (no tm) (Petersen et al., 2011)	0.8235	0.9100	0.6872	0.8667
signalP 4.1 (tm) (Petersen et al., 2011)	0.6471	0.9431	0.6196	0.7951
signalP 3.0 (NN) (Bendtsen et al., 2004)	0.8824	0.9052	0.7220	0.8938
signalP 3.0 (HMM) (Bendtsen et al., 2004)	0.6275	0.9194	0.5553	0.7734
PrediSi (Hiller et al., 2004)	0.3333	**0.9573**	0.3849	0.6453
Philius (Reynolds et al., 2008)	0.6078	0.9336	0.5684	0.7707
Phobius (Käll et al., 2004)	0.6471	0.9289	0.5895	0.7880
signalHsmm-2010	0.9804	0.8720	0.7409	0.9262
signalHsmm-2010 (hom. 50%)	**1.0000**	0.8768	**0.7621**	**0.9384**
signalHsmm-2010 (raw aa)	0.8431	0.9005	0.6853	0.8718
signalHsmm-1987	0.9216	0.8910	0.7271	0.9063
signalHsmm-1987 (hom. 50%)	0.9412	0.8768	0.7194	0.9090
signalHsmm-1987 (raw aa)	0.7647	0.9052	0.6350	0.8350

Conclusions

Thanks to the reduction of amino acid alphabet, signalHsmm is able to recognize signal peptides from the malaria parasites and their relatives more accurately than other software. Simultaneously, our software is still universal enough to provide prediction of other eukaryotic signal peptides on par with the best-performing predictors.

Our model is able to extract decision rules from even very small datasets. Therefore, our model does not need to be permanently retrained with the continuous expansion of sequence databases.

Bibliography

Bendtsen, J. D., Nielsen, H., von Heijne, G., and Brunak, S. (2004). Improved prediction of signal peptides: Signalp 3.0. *Journal of Molecular Biology*, 340(4):783 – 795.

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(suppl 2):W375–W379.

Jain, R. G., Rusch, S. L., and Kendall, D. A. (1994). Signal peptide cleavage regions. functional limits on length and topological implications. *The Journal of Biological Chemistry*, 269(23):16305–16310.

Käll, L., Krogh, A., and Sonnhammer, E. L. L. (2004). A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology*, 338(5):1027–1036.

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.

Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, 8(10):785–786.

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