Predicting eukaryotic signal peptides using hidden Markov models

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

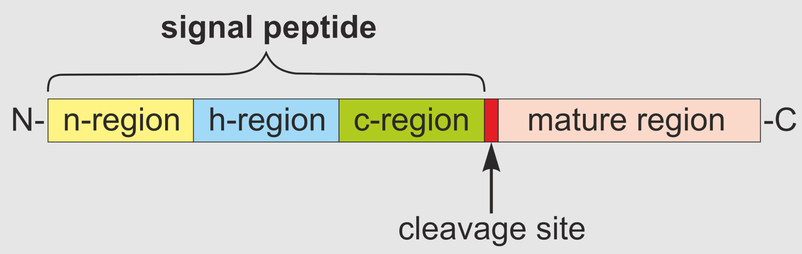
[1]University of Wrocław, Department of Genomics, [2]Wrocław University of Science and Technology, Faculty of Pure and Applied Mathematics

Introduction

Secretory signal peptides:

- are short (20-30 residues) N-terminal amino acid sequences,
- direct a protein to the endomembrane system and next to the extracellular localization,
- possess three distinct domains with variable length and characteristic amino acid composition (Hegde and Bernstein, 2006).
- are universal enough to direct properly proteins in different secretory systems; artifically introduced bacterial signal peptides can guide proteins in mammals (Nagano and Masuda, 2014) and plants (Moeller et al., 2009),
- tag among others hormons, immune system proteins, structural proteins, and metabolic enzymes.

Training of AmyloGram



- 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 the signalHsmm algorithm

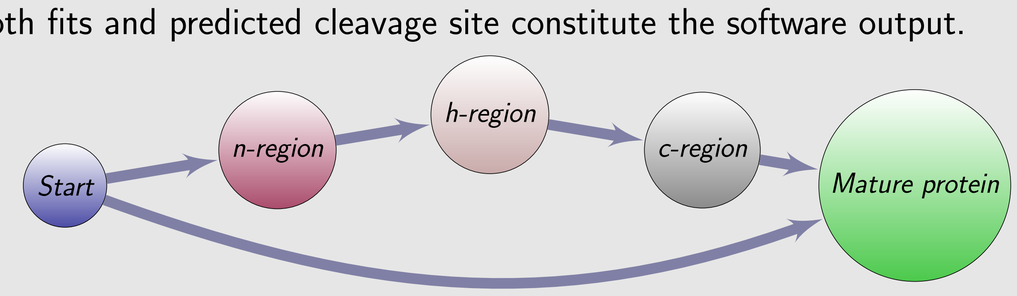
1. Removal of atypical (non-standard amino acids, more than one cleavage site) or poorly annotated records from data set of proteins with signal peptide from UniProtKB 2014_07 (after purification data set contains 3816 eukaryotic proteins with experimentally confirmed signal peptides and 9795 without 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.

Classification of amino acids used by signalHsmm

	Group	Amino acids
Positively charged	1	K, R, H
Nonpolar and aliphatic	2	V, I, L, M, F, W, C
Polar and uncharged	3	S, T, N, Q
Negatively charged and nonpolar	4	D, E, A, P, Y, G

Signal peptide prediction

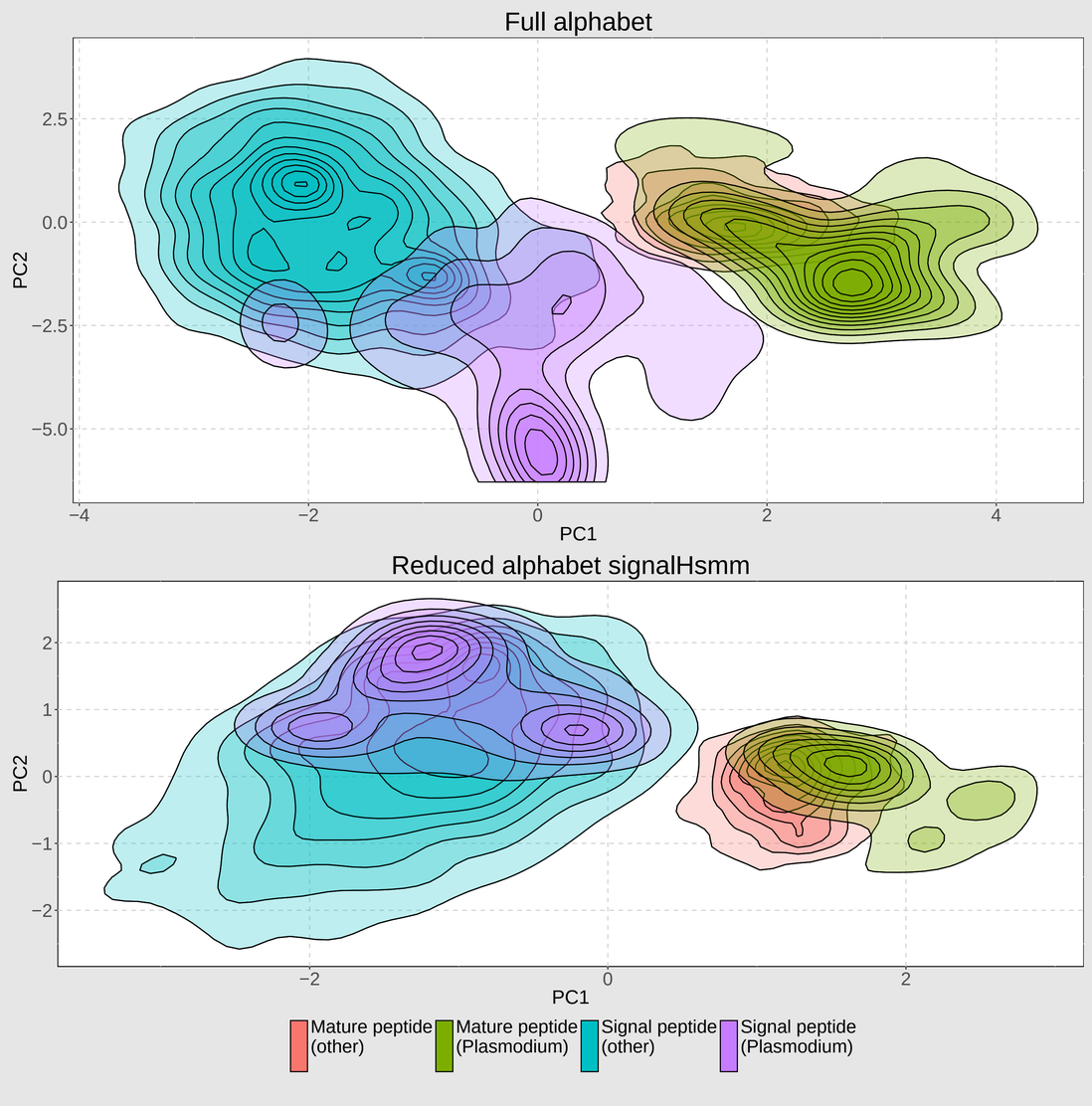
During the test phase, each protein is fitted to two HSMMs representing respectively proteins with and without signal peptides. The probabilities of both fits and predicted cleavage site constitute the software output.



Validation procedure

1. Choose randomly (without replacement) 3816 proteins without signal peptides, reshuffle 3816 proteins with signal peptides.
2. Perform 5-fold cross-validation.
3. Repeat step 1. and 2. 250 times.

Classification of amino acids used by signalHsmm



Comparision with other signal peptide predictors

Benchmark data set: 140 eukaryotic proteins with signal peptide and 280 randomly chosen eukaryotic proteins without signal peptide added after 2010.

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.

Results of comparision

	AUC	H-measure	Gini index	MAM
phobius	0.9643	0.8844	0.9286	1.0809
predsi	0.9411	0.8238	0.8821	0.9302
philius	0.9661	0.8908	0.9321	0.9779
spnotm	0.9679	0.8909	0.9357	0.6739
sptm	0.9750	0.9261	0.9500	0.6889
signalhsmm2010	0.9893	0.8963	0.9786	2.4851
signalhsmm1987	0.9889	0.8994	0.9778	2.4148

MAM - mean absolute cleavage site misplacement.

Conclusions

Hidden semi-Markov models can be used to accurately predict the presence of secretory signal peptides effectively extracting information from very data sets. Prediction of cleavage site position still requires refinement.

Availability and funding

signalHsmm web server:
`www.smorfland.uni.wroc.pl/signalhsmm` qrsignalhsmm.png

This research was partially funded by the KNOW Consortium and National Science Center (2015/17/N/NZ2/01845).

Bibliography

Hegde, R. S. and Bernstein, H. D. (2006). The surprising complexity of signal sequences. *Trends in Biochemical Sciences*, 31(10):563–571.

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.

Moeller, L., Gan, Q., and Wang, K. (2009). A bacterial signal peptide is functional in plants and directs proteins to the secretory pathway. *Journal of Experimental Botany*, 60(12):3337–3352.

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. *Biochemical and Biophysical Research Communications*, 447(4):655 – 659.

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.