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, Poland

[2]Wrocław University of Technology, Institute of Mathematics and Computer Science, Poland

Introduction

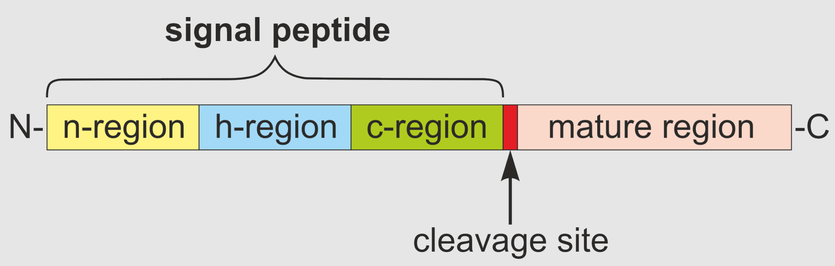
Secretory signal peptides:

- 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).

Proteins with secretory signal peptides are:

- hormons (e.g., prolactin, glucagon),
- immune system proteins (e.g., interferons, interleukins),
- structural proteins (e.g., collagen),
- metabolic enzymes (e.g., alpha-galactosidase, pepsins).

Organization of signal peptide



- 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)

Hidden semi-Markov model of a secretory signal peptide assumes that:

- the observable distribution of amino acids is a result of 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 algorithm

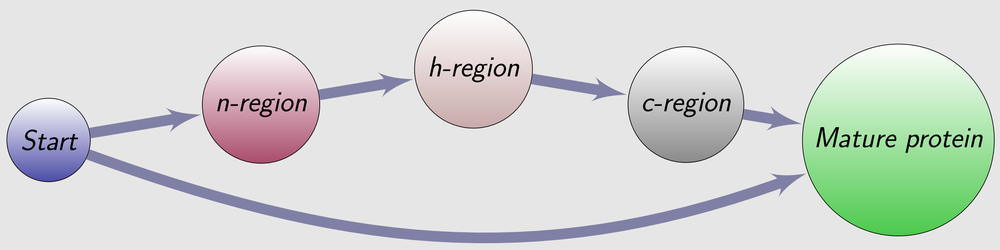
1. Removal of atypical or poorly annotated records from data set of proteins with signal peptide from UniProt database,
2. determination of n-, h-, c-regions by the heuristic algorithm,
3. reduction of dimensionality by aggregating of amino acids to several physicochemical groups,
4. calculation of the 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 signal.hsmm

	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 was fitted to two HSMMs. The outcome consists of probabilities that a particular residue belongs to a given model and predicted cleavage site.



Evaluation

A validated data set contained 3816 eukaryotic proteins with experimentally confirmed signal peptides and 9795 eukaryotic proteins without signal peptides.

Proteins with more than one cleavage site were removed from the data set.

Validation procedure

1. Chose randomly (without replacement) 1200 proteins with signal peptides and train the algorithm (called signal.hsmm).
2. Chose randomly (without replacement) 120 proteins with signal peptides and 120 proteins without signal peptide and test it with a newly trained signal.hsmm. Calculate performance measures.
3. Repeat step 1. and 2. 1000 times.

Results of validation

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.

signal.hsmm1987: trained on data set of 496 eukaryotic proteins with signal peptides added before year 1987.

signal.hsmm2010: trained on data set of 3676 eukaryotic proteins with signal peptides added before year 2010.

Comparision of various software

STH

Summary

Hidden semi-Markov models can be used to accurately predict the presence of secretory signal peptides effectively extracting information from very small data sets.

Avaibility

signal.hsmm web server:

`http://michbur.shinyapps.io/signalhsmm/`

signal.hsmm R package:
`http://cran.r-project.org/web/packages/signal.hsmm/`

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.

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.