

Predicting eukaryotic signal peptides using hidden Markov models

Michał Burdukiewicz^{1*}, Piotr Sobczyk², Paweł Błażej¹, Paweł Mackiewicz¹
*michalburdukiewicz@gmail.com

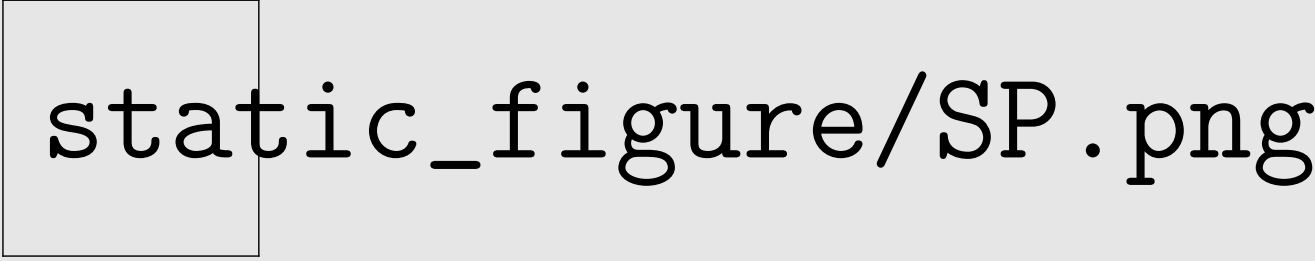
¹University of Wrocław, Department of Genomics, ²Wrocław University of Science and Technology, Faculty of Pure and Applied Mathematics

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 (?).
- are universal enough to direct properly proteins in different secretory systems; artificially introduced bacterial signal peptides can guide proteins in mammals (?) and plants (?),
- tag among others hormones, immune system proteins, structural proteins, and metabolic enzymes.

Training of AmyloGram



- n-region: mostly basic residues (?),
- h-region: strongly hydrophobic residues (?),
- c-region: a few polar, uncharged residues (?).

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 signal.hsmm algorithm

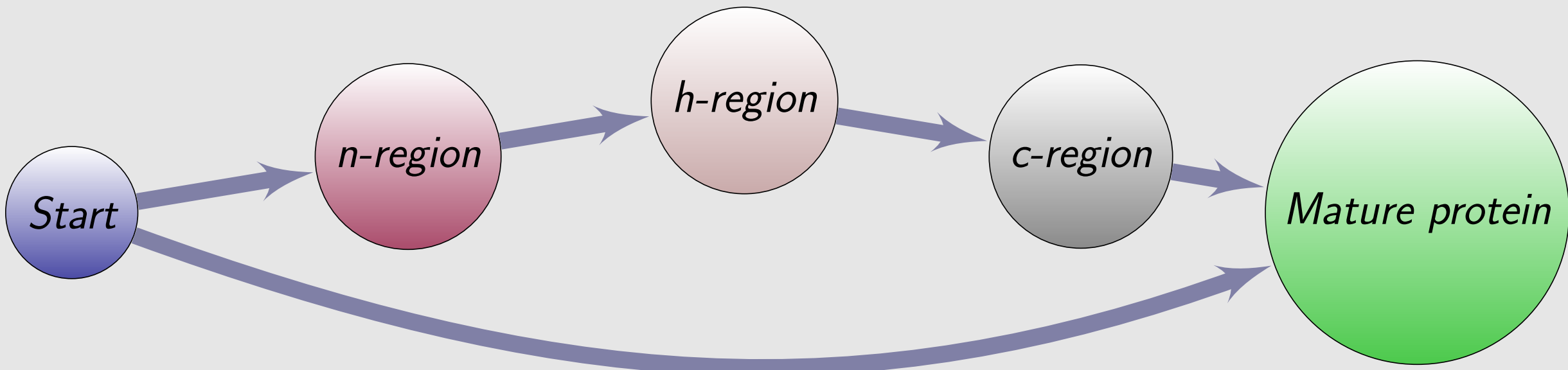
- Removal of atypical (non-standard amino acids, more than one cleavage site) or poorly annotated records from data set of proteins with signal peptide from UniProtKB 2014_07 (after purification data set contains 3816 eukaryotic proteins with experimentally confirmed signal peptides and 9795 without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

Classification of amino acids used by signal.hsmm

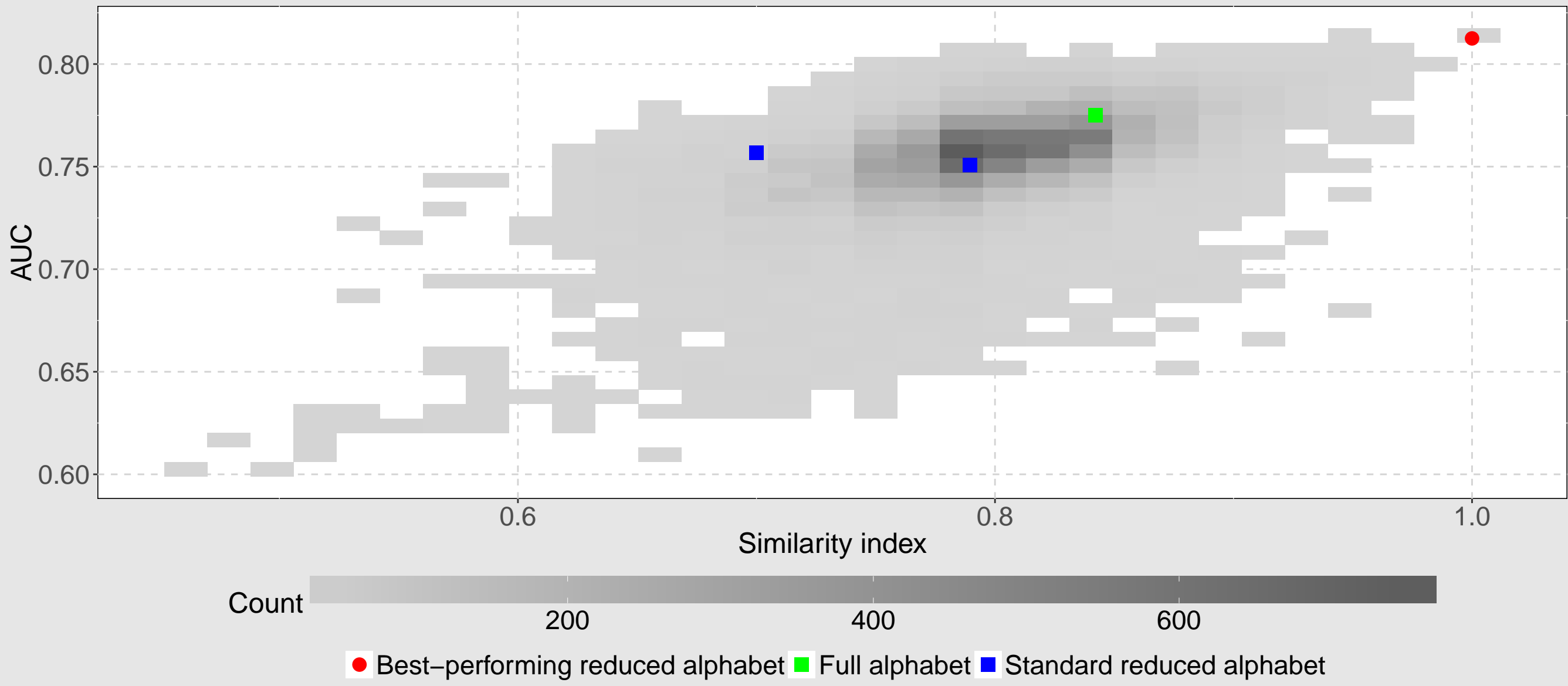
	Group	Amino acids
Positively charged	1	K, R, H
Nonpolar and aliphatic	2	V, I, L, M, F, W, C
Polar and uncharged	3	S, T, N, Q
Negatively charged and nonpolar	4	D, E, A, P, Y, G

Signal peptide prediction

During the test phase, each protein 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.

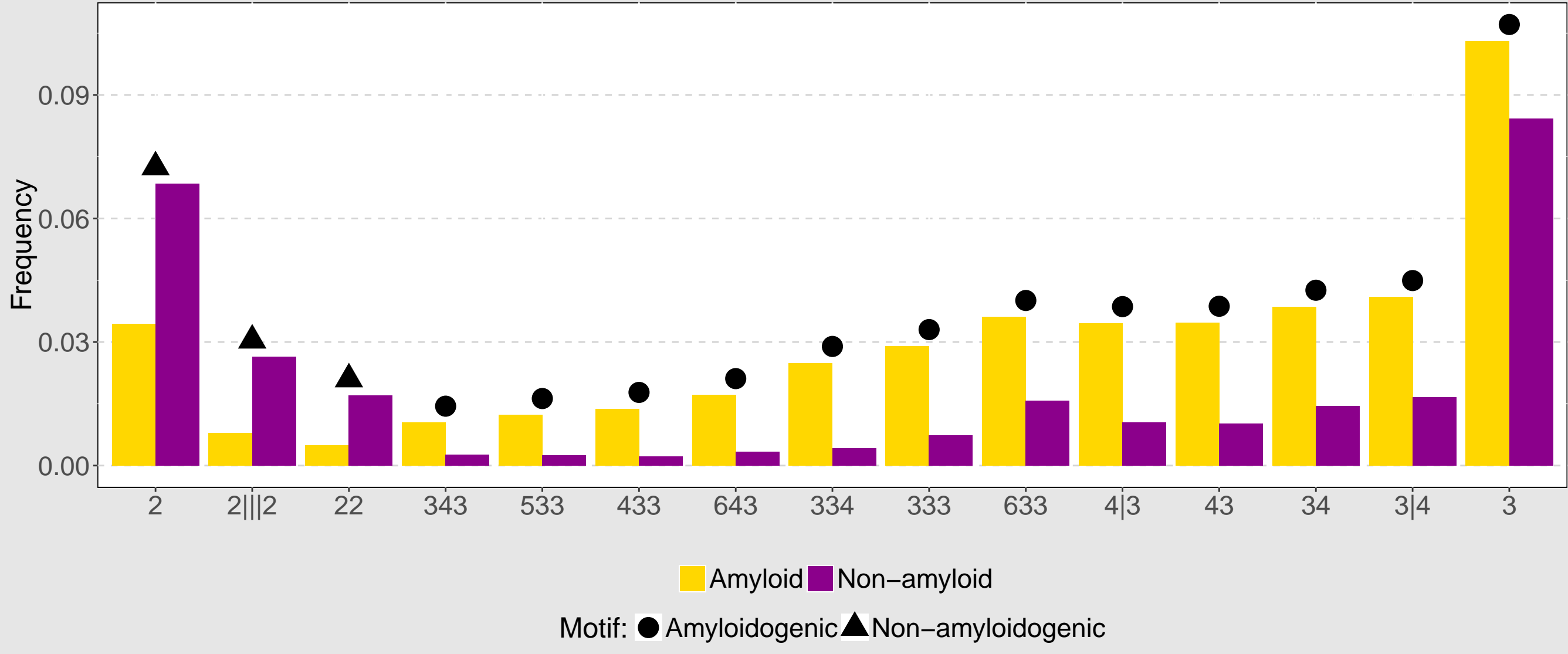


Similarity index



Similarity index (Stephenson and Freeland, 2013) measures the similarity between two reduced alphabets (1 - identical, 0, totally dissimilar). The similarity of a reduced alphabet to the best-performing alphabet is significantly correlated to the AUC of a classifier that employs it. Such relationship indicates that the best-performing reduced alphabet was not found by chance, but represents properties required for the proper prediction of amyloids.

Informative n-grams



The frequency of important n-grams used by the best-performing classifier in amyloid and non-amyloid sequences. The elements of n-grams are amino acids encoded using the best-performing reduced amino acid alphabet. A vertical bar represents a gap in a n-gram between its elements. Dots and triangles denote n-grams occurring in motifs found in respectively amyloidogenic and non-amyloidogenic sequences (Paz and Serrano, 2004).

Benchmark results

Classifier	AUC	MCC	Sensitivity	Specificity
AmyloGram	0.8972	0.6307	0.8658	0.7889
PASTA 2.0(Walsh et al., 2014)	0.8550	0.4291	0.3826	0.9519
FoldAmyloid (Garbuzynskiy et al., 2010)	0.7351	0.4526	0.7517	0.7185
APPNN (Família et al., 2015)	0.8343	0.5823	0.8859	0.7222

The predictor based on the best-performing alphabet, called AmyloGram, was benchmarked against the most popular tools for the detection of amyloid peptides using an external data set *pep424*.

Summary

We identified a group of reduced amino acid alphabets which capture properties of amyloids.

Classifiers based on the full (i.e., unreduced) amino acid alphabet never predicted amyloidogenicity better than the best classifier based on the reduced alphabet.

Our algorithm was also capable of extracting n-gram associated with amyloidogenicity, partially confirming experimental results.

Availability and funding

Our software is available as a web-server:
smorfland.uni.wroc.pl/amylogram.

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

Bibliography

Família, C., Dennison, S. R., Quintas, A., and Phoenix, D. A. (2015). Prediction of Peptide and Protein Propensity for Amyloid Formation. *PLOS ONE*, 10(8):e0134679.

Garbuzynskiy, S. O., Lobanov, M. Y., and Galzitskaya, O. V. (2010). FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics (Oxford, England)*, 26(3):326–332.

Paz, M. L. d. I. and Serrano, L. (2004). Sequence determinants of amyloid fibril formation. *Proceedings of the National Academy of Sciences*, 101(1):87–92.

Stephenson, J. D. and Freeland, S. J. (2013). Unearthing the root of amino acid similarity. *Journal of Molecular Evolution*, 77(4):159–169.

Walsh, I., Seno, F., Tosatto, S. C. E., and Trovato, A. (2014). PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Research*, page gku399.