AmyloGram: prediction of amyloid sequences in R

Jarosław Chilimoniuk1*, Michał Burdukiewicz[2], Piotr Sobczyk[3], Stefan Rödiger[4],
Małgorzata Kotulska[5] and Paweł Mackiewicz[1]

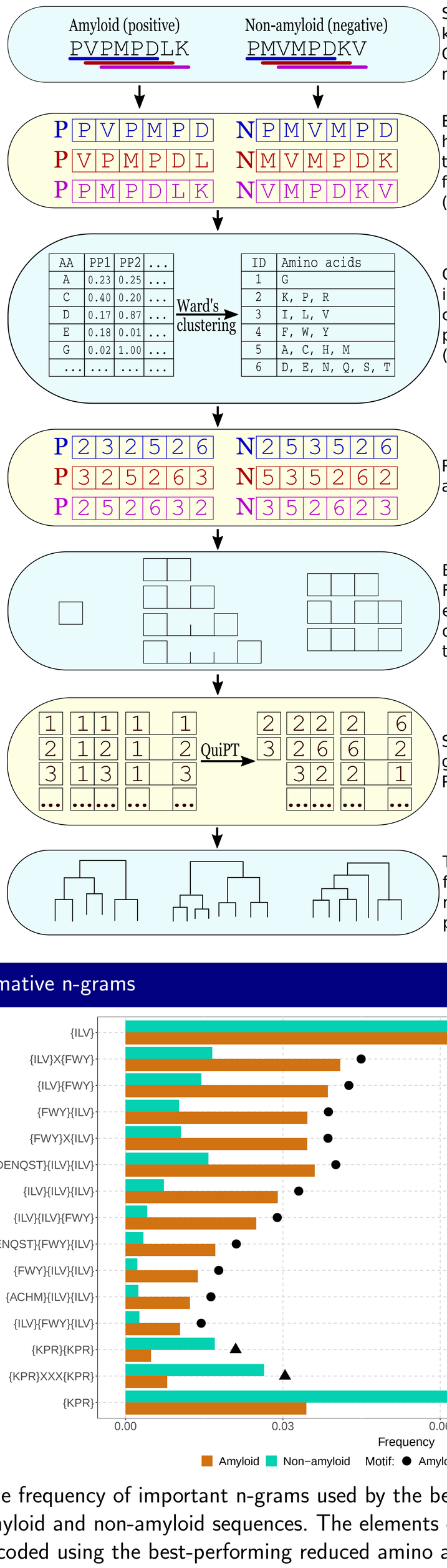
*jaroslaw.chilimoniuk@gmail.com ♢ jarochi@github

[1]University of Wrocław, Department of Bioinformatics and Genomics, [2]Warsaw University of Technology, Faculty of Mathematics and Information Science, [3]Wrocław University of Science and Technology, Faculty of Pure and Applied Mathematics, [4]Brandenburg University of Technology Cottbus-Senftenberg, Institute of Biotechnology, [5]Wrocław University of Science and Technology, Department of Biomedical Engineering

Introduction

Aggregates of amyloid proteins are causes of neurodegenerative disorders. Using features that are easily to interpret, we trained a novel algorithm, AmyloGram, for detection of amyloids. In comparison to other methods predicting amyloids, our software achieved the highest performance. Additionally, we have experimentally confirmed that AmyloGram is able to detect false negatives (amyloid proteins wrongly annotated as non-amyloids) in its training data set.

Scheme



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. X represents any amino acid. Dots and triangles denote n-grams occurring in motifs found in respectively amyloidogenic and non-amyloidogenic sequences (Paz and Serrano, 2004).

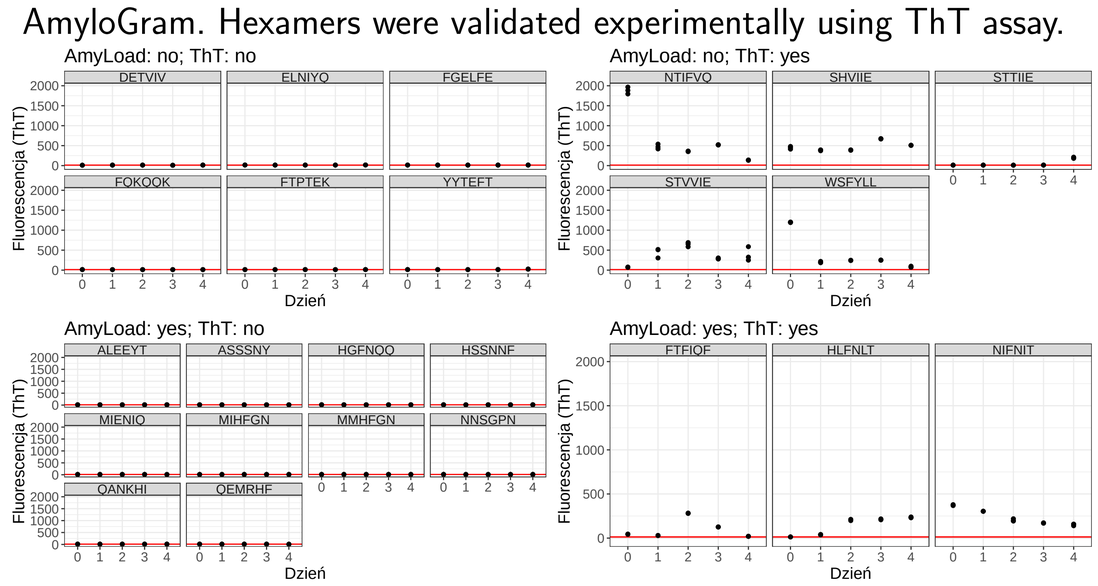
Benchmark results

Classifier	AUC	MCC	Sens.	Spec.
AmyloGram	**0.8972**	**0.6307**	0.8658	0.7889
PASTA (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*.

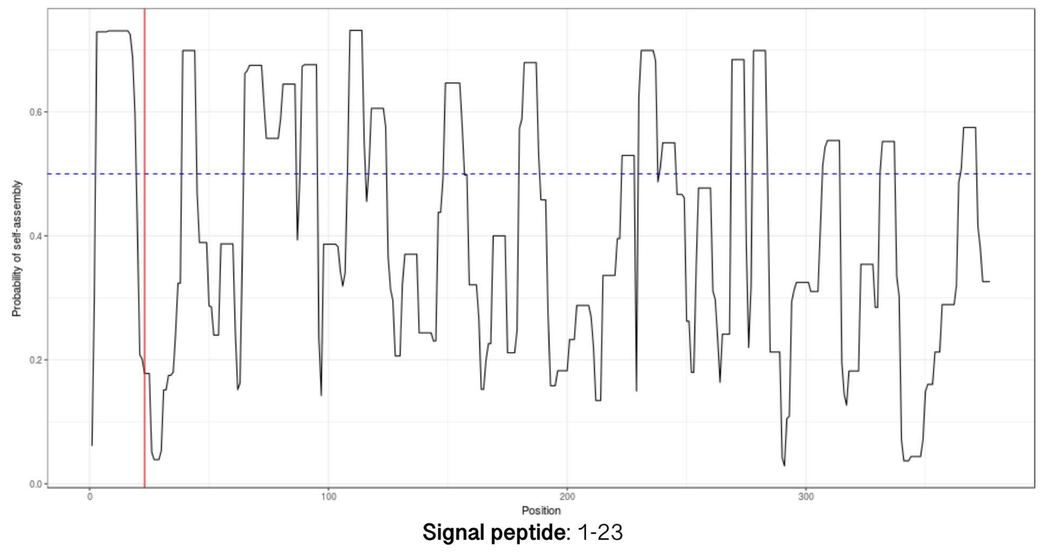
Experimental validation

Using AmyloGram we analyzed all 34 peptides from AmyLoad database. 23 peptides were predicted by AmyloGram otherwise than described in AmyloGram. Hexamers were validated experimentally using ThT assay.



15 out of 23 peptides had the same amyloid properties as predicted by AmyloGram.

MspA - new amyloid protein found by AmyloGram



MspA (WP_011449234.1) has regularly spaced amyloidogenic regions. It's sequence was analysed by Christensen et al. (2018) using AmyloGram (Burdukiewicz et al., 2017). It revealed several regions of predicted high amyloidogenicity (above the horizontal lines). The red vertical line indicates the signal peptide cleavage site position.

Summary and funding

Thanks to the reduction of the amino acid alphabet and description of peptides by short sub-sequences (n-grams), we were able to create the efficient predictor of amyloidogenic sequences called AmyloGram.

Our software is avaible as a web-server:
`www.smorfland.uni.wroc.pl/shiny/AmyloGram/` and **R** package:
`https://cran.r-project.org/package=AmyloGram`.

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

Bibliography

Burdukiewicz, M., Sobczyk, P., Rödiger, S., Duda-Madej, A., Mackiewicz, P., and Kotulska, M. (2017). Amyloidogenic motifs revealed by n-gram analysis. *Scientific Reports*, 7(1):12961.

Christensen, L. F. B., Hansen, L. M., Finster, K., Christiansen, G., Nielsen, P. H., Otzen, D. E., and Dueholm, M. S. (2018). The Sheaths of Methanospirillum Are Made of a New Type of Amyloid Protein. *Frontiers in Microbiology*, 9.

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

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.