AmyloGram: n-gram analysis and prediction of amyloids

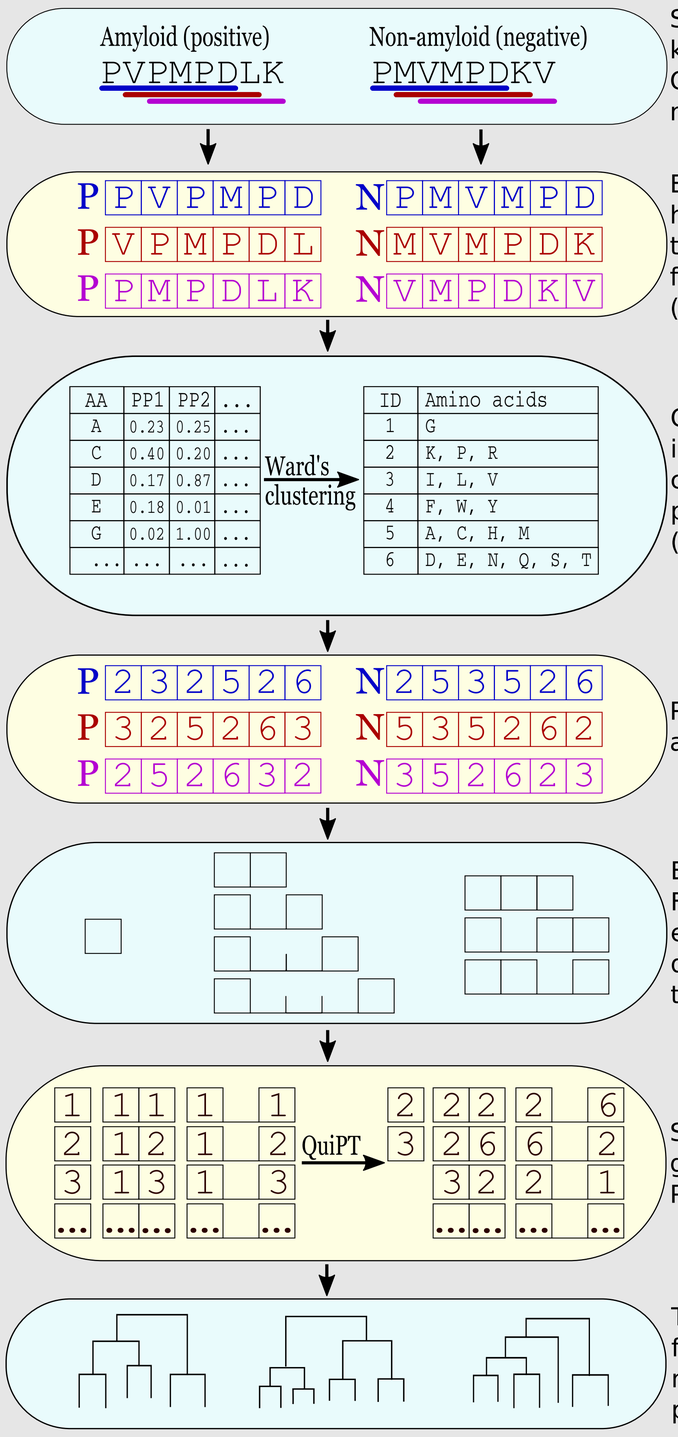
Michał Burdukiewicz[1]*, Piotr Sobczyk[2], Stefan Rödiger[3], Anna Duda-Madej[4], Marlena Gąsior-Głogowska[5], Paweł Mackiewicz[1] and Małgorzata Kotulska[5]
*michalburdukiewicz@gmail.com

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

Introduction

Amyloids are proteins associated with the number of clinical disorders (e.g., Alzheimer's, Creutzfeldt-Jakob's and Huntington's diseases). Despite their diversity, all amyloid proteins can undergo aggregation initiated by 6- to 15-residue segments called hot spots. To find the patterns defining the hot-spots, we trained predictors of amyloidogenicity based on random forests using short subsequences (n-grams) extracted from amyloidogenic and non-amyloidogenic peptides collected in the AmyLoad database.

Scheme



Source data: peptides with known amyloidicity status. Overlapping hexamers are marked by horizontal lines.

Extraction of overlapping hexamers with ascribed the amyloid status taken from their source peptide (P-positive, N - negative).

Clusterization of amino acids into an encoding using a combination of various physicochemical properties (PP).

Reduction of the amino acid alphabet in hexamers.

Extraction of n-grams. From each hexamer, we extracted continuous and discontinous n-grams with the length n = 1, 2 or 3.

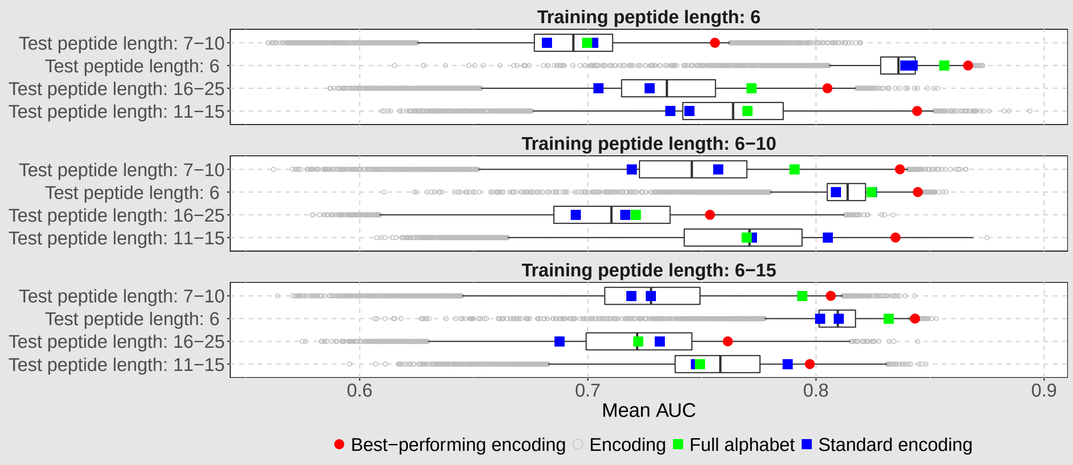
Selection of informative n-grams with Quick Permutation Test (QuiPT).

Training of a random forest classifier using the n-grams selected in the previous step.

Results of cross-validation

The amyloidogenicity of a given peptide may not depend on the exact sequence of amino acids but on its more general properties. Henceforth, we created 524,284 amino acid reduced alphabets (from three to six letters) based on physicochemical properties relevant to amyloidogenicity.

Distribution of mean AUC values of classifiers with various encodings for every possible combination of training and testing data set including different lengths of sequences.



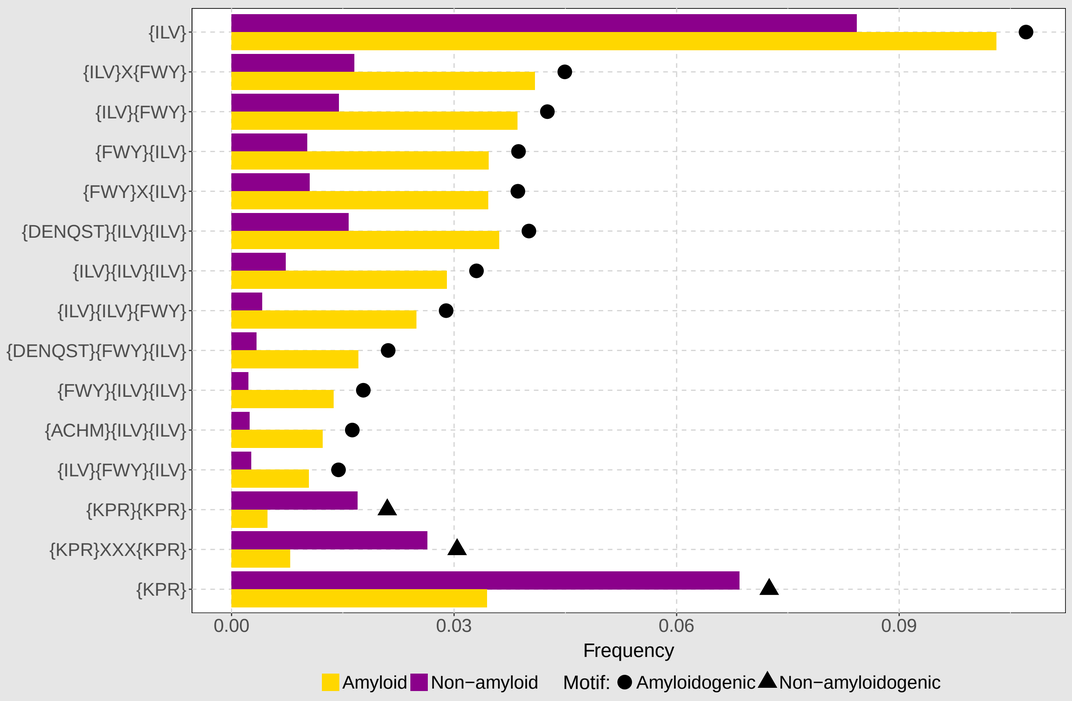
The gray circles correspond to the encodings with the AUC outside the 0.95 confidence interval.

The predictor based on the best-performing encoding reached the highest AUC (0.8667) in classification of the shortest sequences (with the length of 6 residues).

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

The standard encodings found in the literature performed worse than other analyzed encodings in most categories.

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).

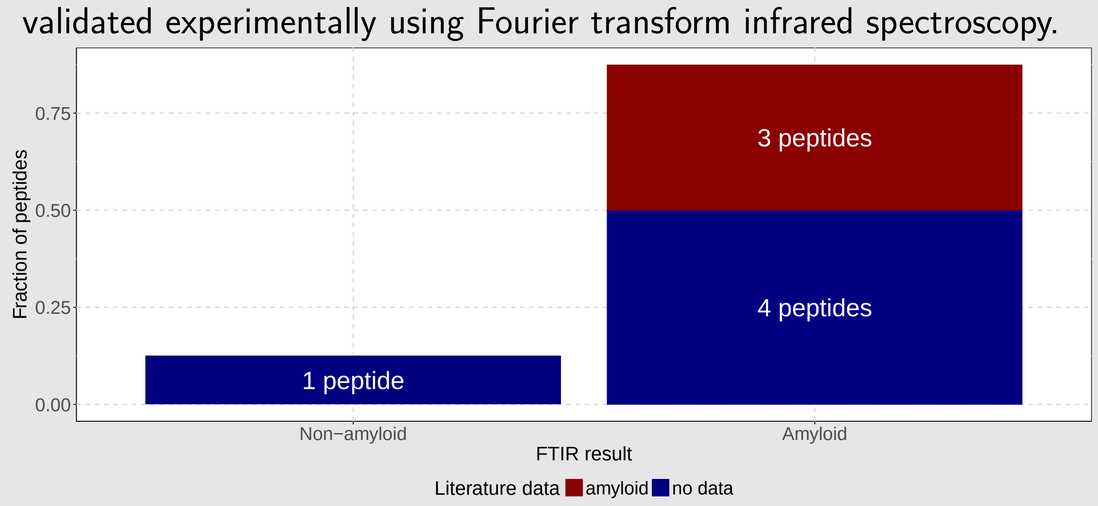
Benchmark results

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 peptides from AmyLoad database. Eight peptides, described in the database as non-amyloids and assesed by AmyloGram with the highest probability of amyloidogenicity, were validated experimentally using Fourier transform infrared spectroscopy.



Seven out of eight peptides had amyloidogenic properties. In addition, three of them were annotated as amyloids by other research groups.

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

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.

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.