AmyloGram: analysis and prediction of amyloids using n-grams

Michał Burdukiewicz[1], Piotr Sobczyk[2], Stefan Rödiger[3], Paweł Mackiewicz[1] and Małgorzata Kotulska[4]
*michalburdukiewicz@gmail.com

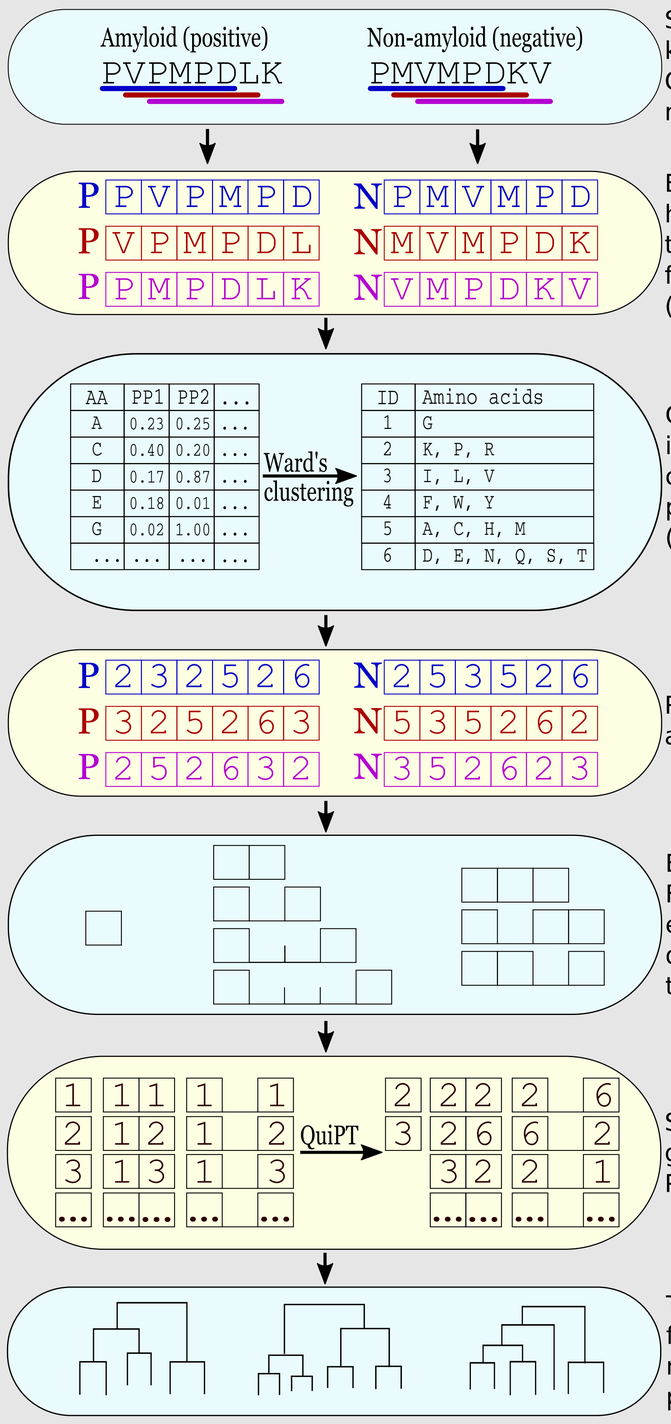
[1]University of Wrocław, Department of Genomics, [2]Wrocław University of Science and Technology, Faculty of Pure and Applied Mathematics, [3]Brandenburg University of Technology Cottbus-Senftenberg, Institute of Biotechnology, [4]Wrocław University of Science and Technology, Department of Biomedical 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 multiple predictors of amyloidogenicity based on random forests using short subsequences (n-grams) extracted from amyloidogenic and non-amyloidogenic peptides collected in the AmyLoad database.

The best-performing predictor, AmyloGram, was compared to state-of-art predictors of amyloids using the independent benchmark dataset.

Training of AmyloGram

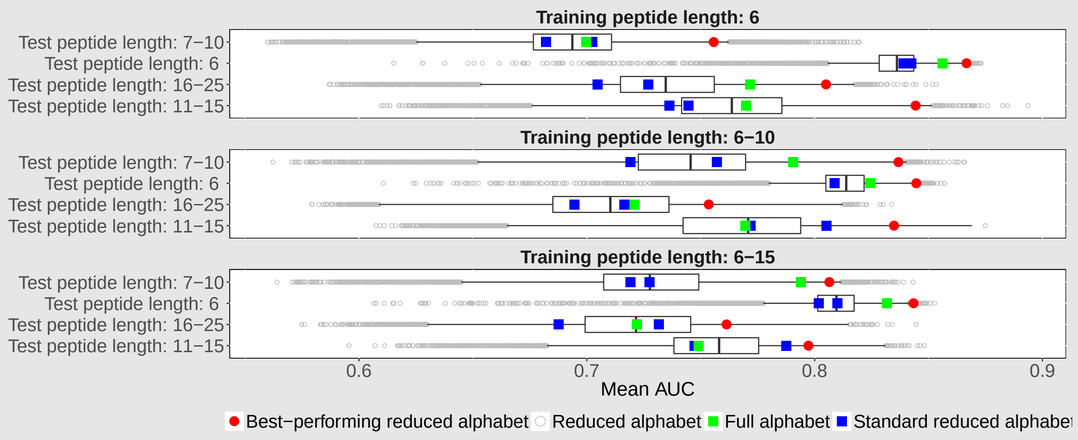


Reduced amino acid alphabet

The amyloidogenicity of a given peptide may not depend on the exact sequence of amino acids but on its more general properties. We handpicked 17 measures from AAIndex database describing features important in the amyloidogenicity, such as: size of residues, hydrophobicity, solvent surface area, frequency in β-sheets and contactivity. Based on that we created 524,284 amino acid reduced alphabets with different level of amino acid alphabet reduction from three to six amino acid groups.

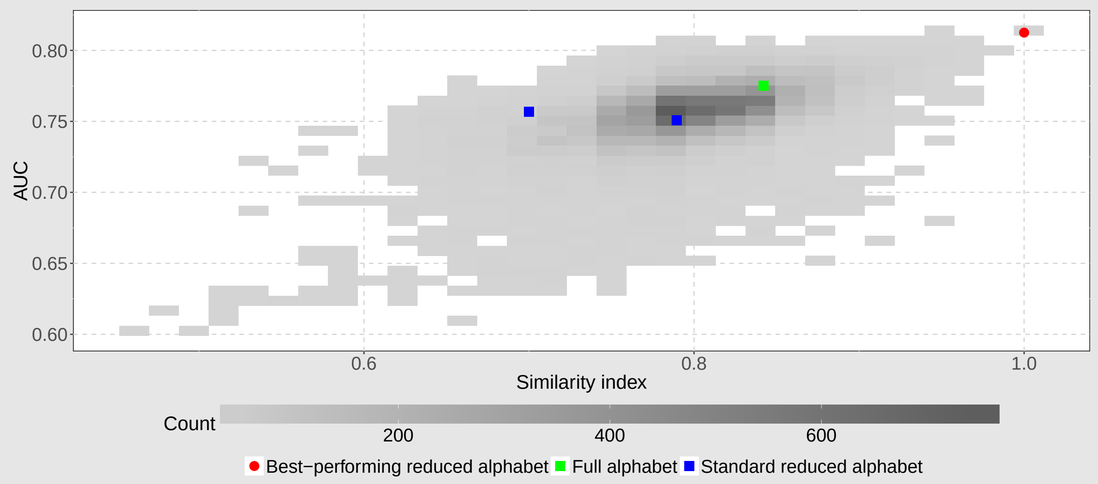
Results of cross-validation

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



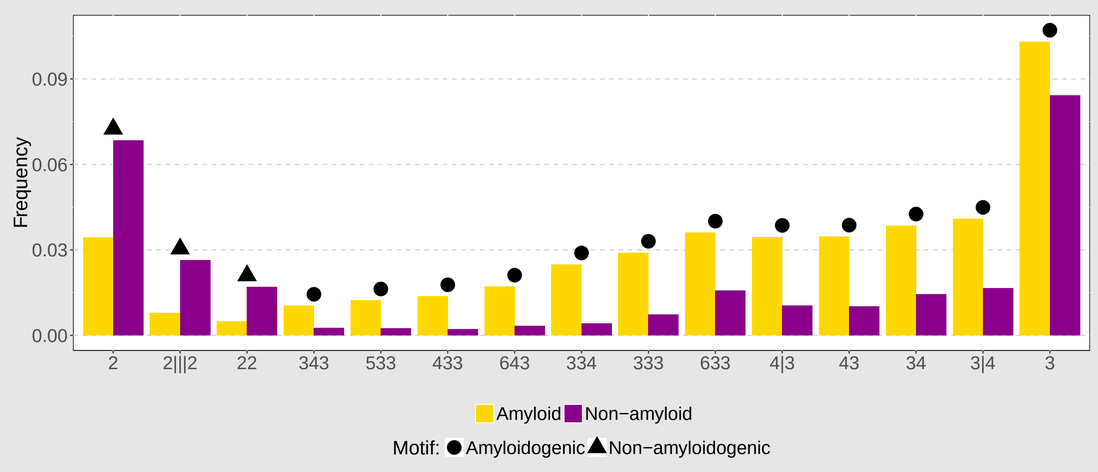
The left and right hinges of boxes correspond to the 0.25 and 0.75 quartiles. The bar inside the box represents the median. The gray circles correspond to the reduced alphabets with the AUC outside the 0.95 confidence interval.

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.