AmyloGram: n-gram analysis and prediction of amyloids

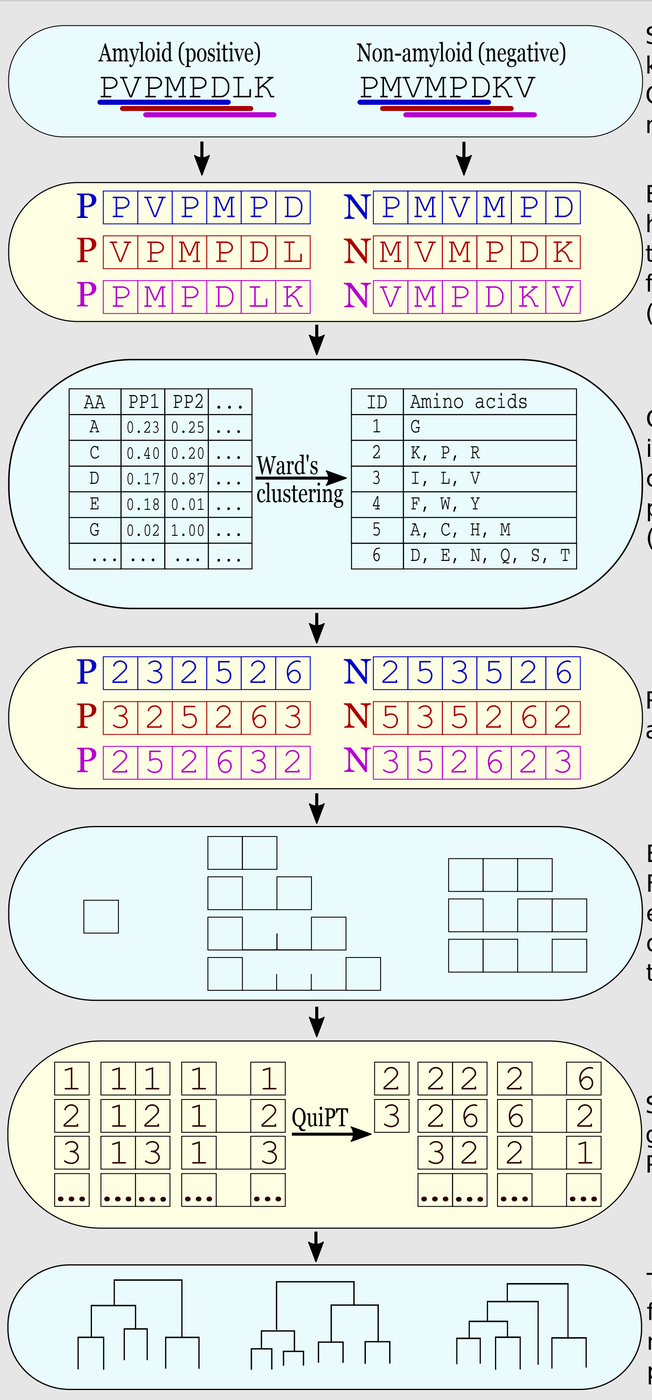
Michał Burdukiewicz[1], Piotr Sobczyk[2], Paweł Mackiewicz[1] and Małgorzata Kotulska[3]
*michalburdukiewicz@gmail.com

[1]University of Wrocław, Department of Genomics
[2]Wrocław University of Science and Technology, Faculty of Pure and Applied Mathematics
[3]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.

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 data base 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 encodings with different level of amino acid alphabet reduction from three to six amino acid groups using Ward's clusterization (Ward, 1963), which was performed on all combinations of the normalized values of physicochemical properties from 1 to 17.

Quick Permutation Test (QuiPT)

Model and statistic independent permutation tests can be used to filter features obtained through counting n-grams.
During a permutation test class labels are randomly exchanged during computation of a significance statistic. p-values are defined as:

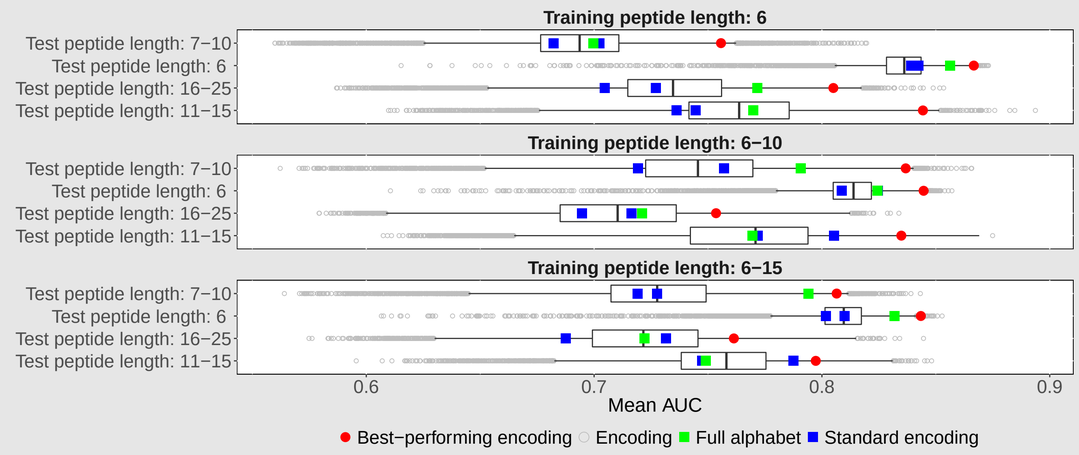
$$\text{p-value} = \frac{N_{T_P > T_R}}{N}$$

where $N_{T_P > T_R}$ is number of times when T_P (permuted test statistic) was more extreme than T_R (test statistic for non-permuted data).
Permutation tests are computationally expensive (especially considering precise estimation of small p-values, because the number of permutations is inversely proportional to the interval between p-values).

Quick **P**ermutation **T**est (QuiPT) thanks to the unique parameterization replaces a permutation test with the exact two-sided Fisher's test (Lehmann, 1986) reducing the computation cost.

Results of cross-validation

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.



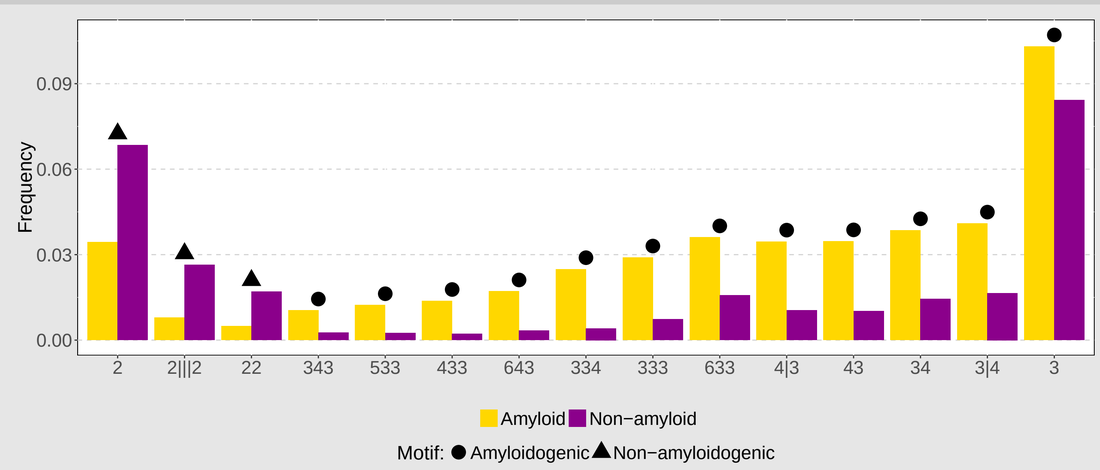
The left and right hinges of boxes correspond to the 0.25 and 0.75 quartiles. The bar inside the box represents the median. The gray circles correspond to the encodings with the AUC outside the 0.95 confidence interval.

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. 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 (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 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:
`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.

Lehmann, E. (1986). *Testing statistical hypotheses*. Wiley series in probability and mathematical statistics: Probability and mathematical statistics. Wiley.

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.

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244.