N-gram analysis of amyloid data

Michał Burdukiewicz[1], Piotr Sobczyk[2], Paweł Mackiewicz[1] and Małgorzata Kotulska[3]

[1]University of Wrocław, Department of Genomics, Poland
[2]Wrocław University of Technology, Department of Mathematics, Poland
[3]Wrocław University of Technology, Department of Biomedical Engineering, Poland

Aim

Investigate features responsible for amyloidogenicity, the cause of various clinical disorders (e.g. Alzheimer's or Creutzfeldt-Jakob's diseases). The features are defines as countinous and discontinous subsequences of amino acids (n-grams).

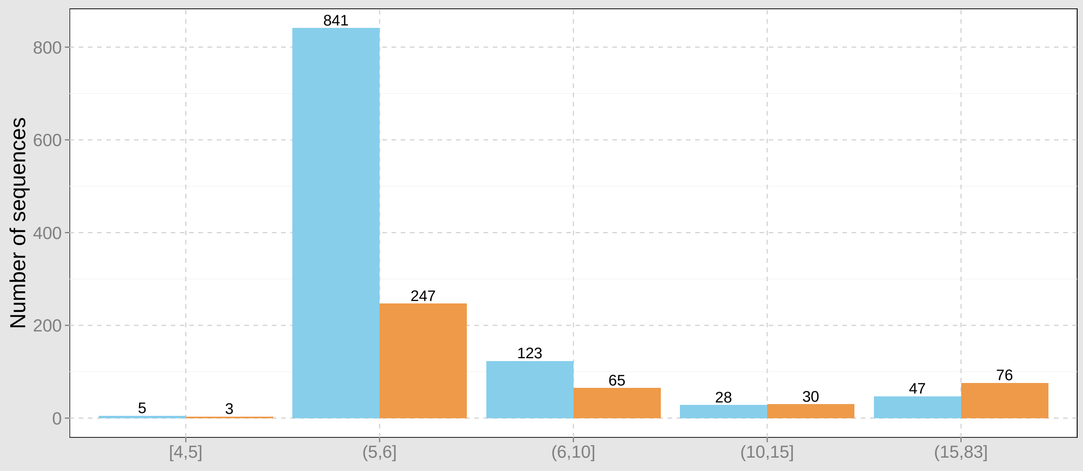
Introduction

All amyloidogenic sequences (amyloids), despite their variability in size and amino acid composition, form mostly cytotoxic aggregates (Breydo and Uversky, 2015). The hallmark trait of amyloids is the presence of hot-spots, short sequences of amino acids that play key role in the aggregation process (Fändrich, 2012).

The n-gram encoding of sequences creates high-dimensional data sets. We filtered significant features using the **Qui**ck **P**ermutation **T**est (**QuiPT**) and information gain criterion with significance level $\mathbf{0.95}$ (Burdukiewicz et al., 2015).

AmyLoad database

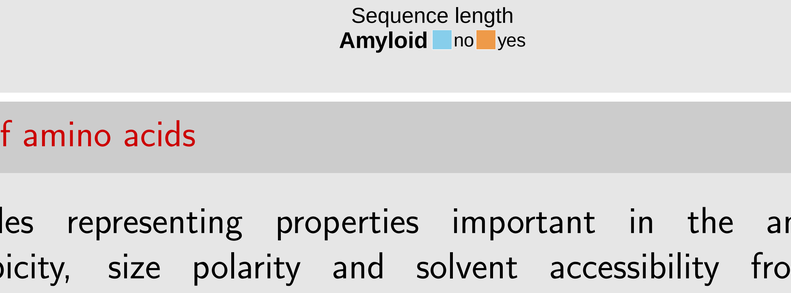
The sequences used in the study (1044 non-amyloids and 421 amyloids) were extracted from AmyLoad database (Wozniak and Kotulska, 2015).



Clustering of amino acids

1. Nine scales representing properties important in the amylogenicity: hydrophobicity, size polarity and solvent accessibility from AAIndex database (Kawashima et al., 2008) were chosen. Additionally, two frequencies of forming contact sites (Wozniak and Kotulska, 2014) were added. All scales were normalized.
2. All combinations of characteristics (each time selecting only one scale per the property) were clustered using Euclidean distance and Ward's method.
3. Each clustering was divided into 3 to 6 groups creating 144 encodings of amino acids. Redundant 51 encodings (identical to other encodings) were removed.

Evaluation



1. Sequences shorter than 6 amino acids were discarded.
2. From each sequence overlapping windows of length 6 were extracted. All windows were labelled as their sequence of the origin, e.g. all windows extracted from amyloid sequence were labelled as positive (see Figure A and B).
3. For each window, 1-, 2- and 3-grams (both discontinous and continous) were extracted (see Figure B). For each encoding, the encoded n-grams were filtered by the QuiPT and used to train the Random Forests (Liaw and Wiener, 2002). This procedure was performed independently on three training sets: a) 6 amino acids, b) 10 amino acids or shorter, c) 15 amino acids or shorter creating three classifiers.
4. All classifiers were evaluated in the 5-fold cross-validation eight times. The sequence was labelled as positive (amylogenic), if at least one window was assessed as amylogenic.

Cross-validation results

Encodings with the best sensitivity and specificity for each training set type.

Training length	Number of groups	Encoding ID	Sensitivity	Specificity
6	4	45	0.5038	0.9014
<16	6	87	0.9195	0.5186

The best encodings

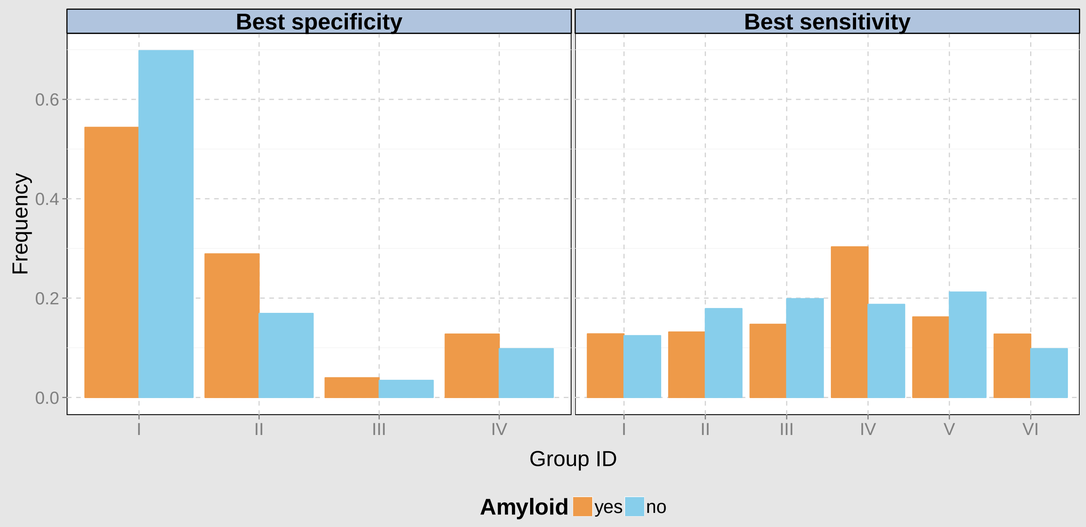
The best specificity encoding.

ID	Amino acids
I	H, M
II	F, W, Y
III	C, I, L, V
IV	A, D, E, G, K, N, P, Q, R, S, T

The best sensitivity encoding.

ID	Amino acids
I	A, T
II	D, E, N
III	G, P, S
IV	F, W, Y
V	H, K, Q, R
VI	C, I, L, M, V

All 1-grams were considered as significant features by QuiPT.



Considering the hydrophobic 1-grams, only aromatic residues differ amyloid and non-amyloid sequences.

AmyloGram

The best specificity encoding (training sequence maximum length 6, 4 groups) and the best sensitivity (training sequence maximum length <16, 6 groups) seem to have the different areas of the competence.
AmyloGram, the committee of the best specificity and best sensitivity classifiers, has overall $\mathbf{0.8911}$ AUC, $\mathbf{0.7473}$ sensitivity and $\mathbf{0.8684}$ specificity.

Comparison with other software

We used established *pep424* benchmark data set (Walsh et al., 2014) to compare AmyloGram with top-performing predictors of amyloidogenicity (default settings used).

Results of benchmark.

Predictor name	AUC	Sensitivity	Specificity
AmyloGram	0.8426	0.8054	0.7222
PASTA2	0.7920	0.7248	0.8593
FoldAmyloid	0.7351	0.7517	0.7185

Since AmyloGram model assumes 6 amino acids as the minimum length of amyloid sequence, the five short sequences ($\mathbf{1.18\%}$) were removed from data set.

Summary and availability

AmyloGram is a model-independent predictor of amylogenicity. Instead, it provides insight on the structural features present in the hot-spots. Moreover, AmyloGram recognises amylogenic sequences better than existing predictors.

AmyloGram web-server: `smorfland.uni.wroc.pl/amylogram`.

Bibliography

Breydo, L. and Uversky, V. N. (2015). Structural, morphological, and functional diversity of amyloid oligomers. *FEBS letters*.

Burdukiewicz, M., Sobczyk, P., and Lauber, C. (2015). *biogram: analysis of biological sequences using n-grams*. R package version 1.2.

Fändrich, M. (2012). Oligomeric Intermediates in Amyloid Formation: Structure Determination and Mechanisms of Toxicity. *Journal of Molecular Biology*, 421(4–5):427–440.

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. *Nucleic Acids Research*, 36(suppl 1):D202–D205.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 2(3):18–22.

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*, 42(W1):W301–W307.

Wozniak, P. P. and Kotulska, M. (2014). Characteristics of protein residue-residue contacts and their application in contact prediction. *Journal of Molecular Modeling*, 20(11).

Wozniak, P. P. and Kotulska, M. (2015). AmyLoad: website dedicated to amyloidogenic protein fragments. *Bioinformatics (Oxford, England)*.