# 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

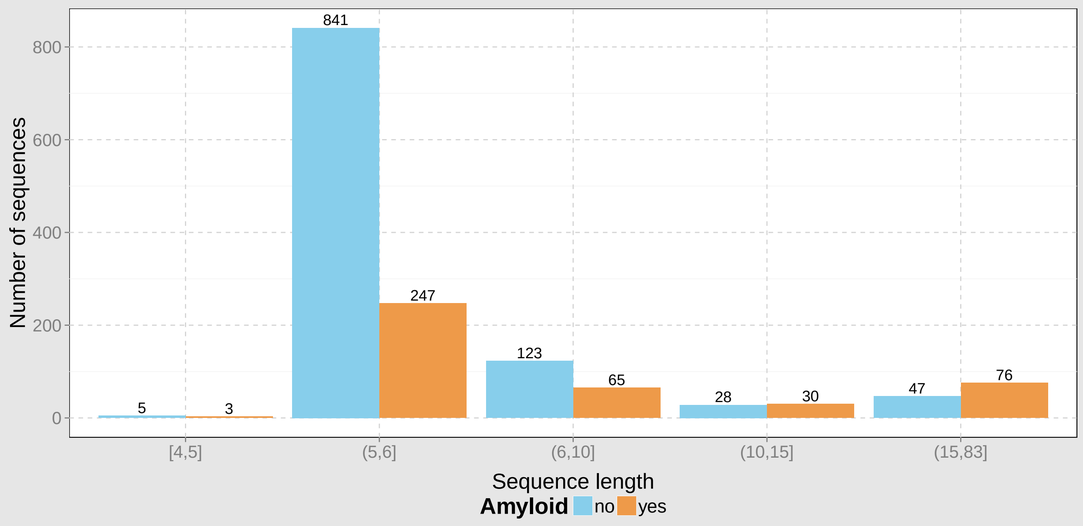
Investigate and predict sequences responsible for amyloidogenicity, the cause of various clinical disorders (e.g. Alzheimer's or Creutzfeldt-Jakob's diseases).

## 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).

## AmyLoad database

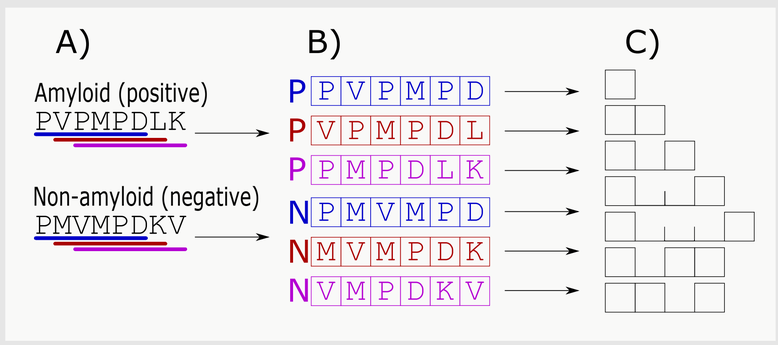
The sequences used in the study (1044 non-amyloids and 421 amyloids) were extracted from AmyLoad database (Wozniak and Kotulska, 2015).




## 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.
4. 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.

## Encoding distance

The encoding distance between **A** and **B** is defined as the minimum number of amino acids that have to be shifted between subgroups of encoding **A** to make it identical to **B** (order of subgroups in the encoding and amino acids in a group is unimportant).

| Group | Elements |
|---|---|

| Group | Elements |
|---|---|

## Specificity versus sensitivity

| Training length | Number of groups | Encoding ID | AUC | Specificity | Sensitivity |
|---|---|---|---|---|---|
| 6 | 3 | 6 | 0.7955 | 0.8221 | 0.6181 |
| 6 | 4 | 45 | 0.8183 | 0.9014 | 0.5038 |
| <11 | 4 | 2 | 0.6615 | 0.4304 | 0.8307 |
| <11 | 3 | 15 | 0.8088 | 0.8329 | 0.6060 |
| <16 | 3 | 16 | 0.8162 | 0.7477 | 0.7374 |
| <16 | 6 | 87 | 0.8320 | 0.5186 | 0.9195 |

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

The best specificity encoding (training length 6, 4 groups, encoding ID 45) and the best sensitivity (training length <16, 6 groups, encoding ID 87) seem to have the different areas of the competence.
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.

## Bibliography

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

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.

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)*.