N-gram analysis of amyloid data

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

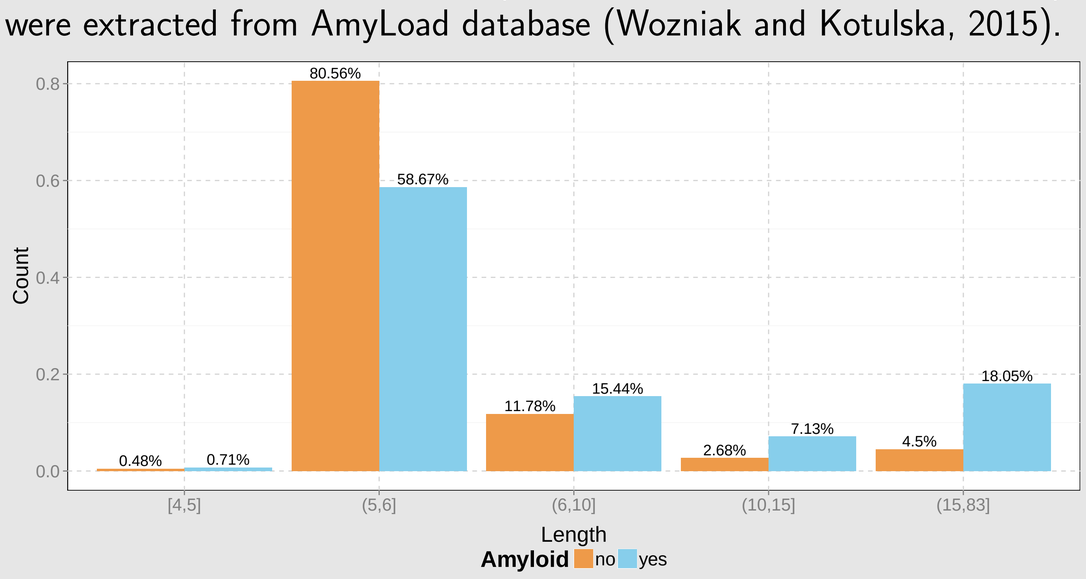
[1]University of Wrocław, Department of Genomics, Poland
[2]Wrocław University of Technology, Department of Mathematics, Poland
[3]Wrocław University of Technology, Department of Biomedical Engineering, Poland

Aim

Investigate putative motifs of hot-spots, short sequences responsible for amyloidogenicity which is associated with a number of clinical disorders, for example Alzheimer's or Creutzfeldt-Jakob's diseases.

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. Scales were normalized between 0 i 1.
2. All combinations of characteristics (only one scale per 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 5-grams were extracted. All n-grams were labelled as their sequence of the origin (e.g. 5-grams extracted from amyloid sequence were labelled as positive).
3. For each encoding features were filtered by the QuiPT and used to train the Random Forests (Liaw and Wiener, 2002). The procedure was performed on three traning 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. The sequence was labelled as positive if at least one 5-gram was assessed as amylogenic.
5. The cross-validation was repeated 8 times.

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
1	a, b, c
2	d, e

Encoding **A**.

Group	Elements
1	a, b
2	d, e
3	c

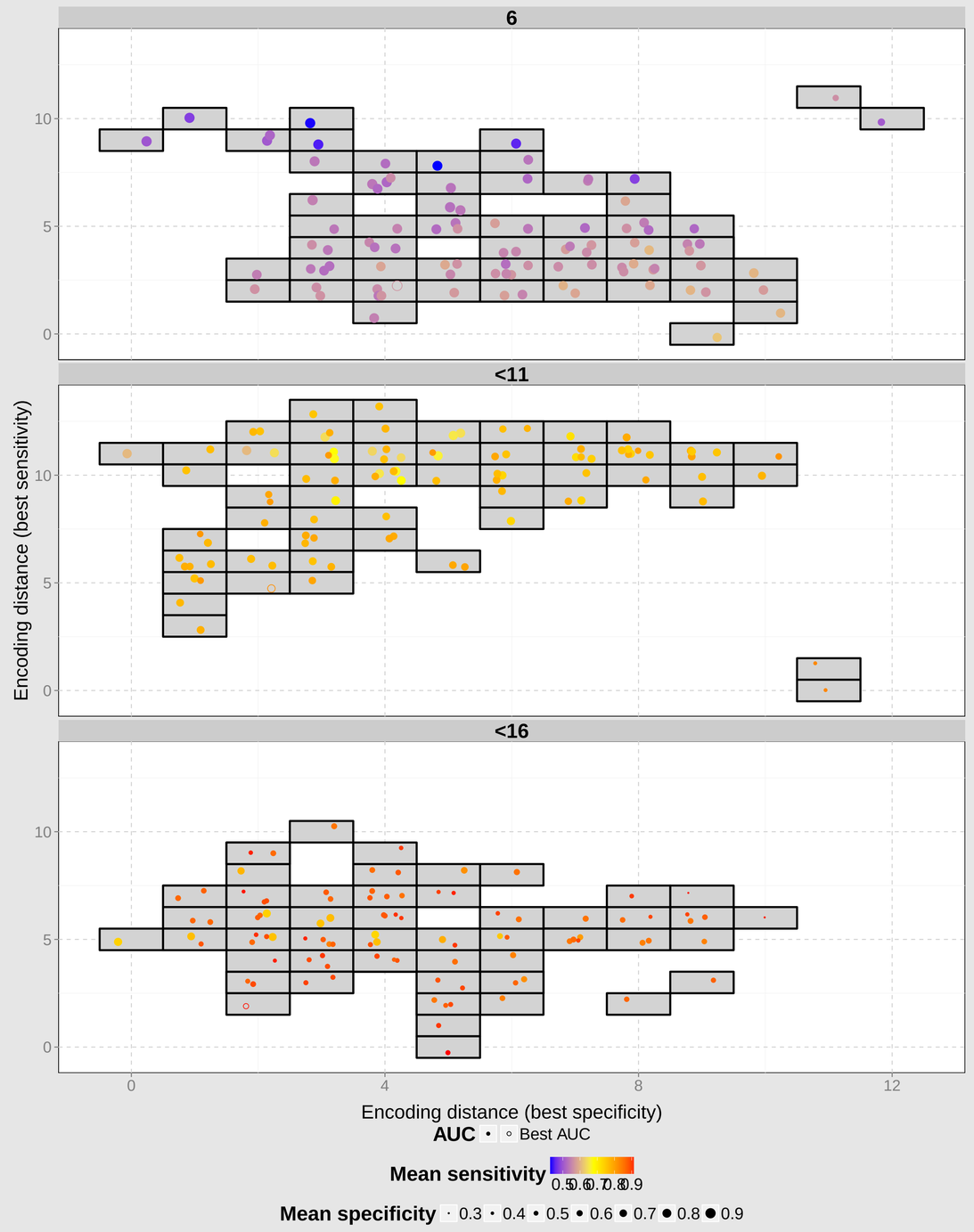
Encoding **B**.

The encoding distance between **A** and **B** is 1 (element *c* must be moved from Group 3 to Group 1).

Specificity versus sensitivity

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

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



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)*.