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

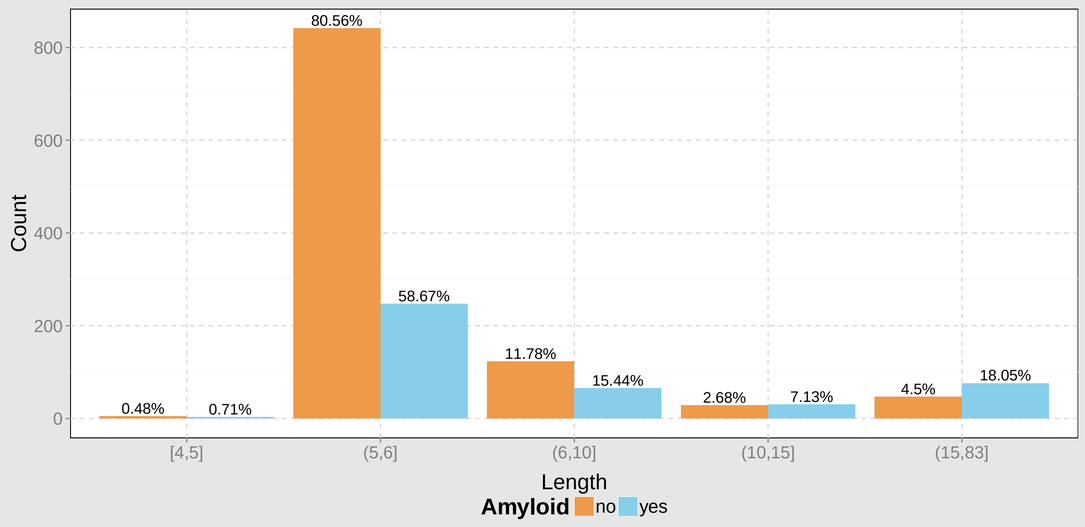
[3]Wrocław University of Technology, Department of Biomedical Engineering, Poland

Introduction

Stuff

AmyLoad database

The sequences used in the study were extracted from AmyLoad database (Wozniak and Kotulska, 2015).



The height of the bar is equal to the number of sequences within given length interval in the database. The percentage above the bar represent fraction sequences within given length interval among amyloid and non-amyloid sequences.

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.
2. Scales were normalized between 0 i 1.
3. All combinations of characteristics (only one scale per property) were clustered using Euclidean distance and Ward's method.
4. Each clustering was divided into 3 to 6 groups creating 144 encodings of amino acids.
5. Redundant 51 encodings (identical to other encodings) were removed.

Encoding distance

The encoding distance between **A** and **B** is defined as the minimum number of amino acids that have to be moved between subgroups of encoding to make **A** 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, c

Encoding **B**.

The encoding distance between **A** and **B** is 1 (element *c* must be moved from Group 2 to Group 1).

Evaluation

1. Sequences shorter than 6 amino acids were discarded.
2. From each sequence 5-grams were extracted. In case of positive sequences, all extracted 5-grams were labelled as positive. In case of negative sequences, all extracted 5-grams were marked as negative.
3. For each encoding three Random Forests (Liaw and Wiener, 2002) were trained on sequences with length respectively a) 6 amino acids, b) 10 amino acids or shorter, c) 15 amino acids or shorter.
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.

More stuff

Stuff

Bibliography

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)*.