N-gram analysis of amyloid data

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 Technology, Faculty of Pure and Applied Mathematics
[3]Wrocław University of Technology, Department of Biomedical Engineering

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).

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.

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.

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`.

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.

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).