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, Faculty of Fundamental Problems of Technology, Poland*
malgorzata.kotulska@pwr.edu.pl

Amyloids are short proteins associated with a number of clinical disorders, for example Alzheimer's or Creutzfeldt-Jakob's diseases. Despite their variability in size and amino acid composition, amyloidogenic sequences form mostly cytotoxic aggregates, although a few of them is biologically functional [BU15]. The hallmark trait of amyloids is the presence of characteristic short sequences of amino acids, called hot-spots. Furthermore, amyloids can create zipper-like β-structures [F12]. Although studies investigating properties of amyloidogenic sequences have already been conducted, the newly established AmyLoad database facilities large-scale analysis of amyloids [WK15].

Among commonly acclaimed methods of predicting amyloids, FISH Amyloid [GK14] focuses more on the putative motifs of hot spots. To expand its model by considering longer and more complicated motifs, we used n-gram analysis. N-grams (k-mers) are vectors of n characters derived from input sequences. The number of possible n-grams is equal to u^n, where u is the length of the alphabet (4 in case of nucleic acids and 20 in case of proteins). To deal with the dimensionality of the problem, we implemented QuiPT (Quick Permutation Test) in *biogram* software [BSL15], which performs an exact test instead of a large number of permutations.

To reduce the dimension even more, we grouped amino acids into clusters based on their physicochemical properties potentially important in the amyloid type of aggregation. The features include several scales quantitatively representing hydrophobicity, size and accessibility derived from AAIndex database, and propensity of amino acids to form contact sites derived in [WK14].

The n-gram model, trained on the data from AmyLoad database, is validated through amyloid prediction framework using random forests. The preliminary analysis of the amyloidogenic sequences not only facilitates prediction of amyloids but also gives a new insight into the physicochemical characteristics of the hot spots. The mean AUC of the classifier committee in 5-fold cross-validation was 0.89. The most balanced classifier, regarding its sensitivity S_n and specificity S_p, enabled the predictions with $S_n = 0.75$, $S_p = 0.87$, and AUC = 0.89.

The address of amylogenicity predictor (AmyloGram): www.smorfland.uni.wroc.pl/amylogram.

References

[BSL15] Michal Burdukiewicz, Piotr Sobczyk, and Chris Lauber. *biogram: analysis of biological sequences using n-grams*. 2015. R package version 1.2.

[BU15] Leonid Breydo and Vladimir N. Uversky. Structural, morphological, and functional diversity of amyloid oligomers. *FEBS letters*, July 2015.

[F12] Marcus Fndrich. Oligomeric Intermediates in Amyloid Formation: Structure Determination and Mechanisms of Toxicity. *Journal of Molecular Biology*, 421(45):427–440, August 2012.

[GK14] Pawel Gasior and Malgorzata Kotulska. FISH Amyloid a new method for finding amyloidogenic segments in proteins based on site specific co-occurence of aminoacids. *BMC Bioinformatics*, 15(1):54, February 2014.

[WK14] Pawel P. Wozniak and Malgorzata Kotulska. Characteristics of protein residue-residue contacts and their application in contact prediction. *Journal of Molecular Modeling*, 20(11), 2014.

[WK15] Pawel P. Wozniak and Malgorzata Kotulska. AmyLoad: website dedicated to amyloidogenic protein fragments. *Bioinformatics (Oxford, England)*, June 2015.