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*
malgorzata.kotulska@pwr.edu.pl

Amyloids are short proteins associated with the number of clinical disorders, for example Alzheimer's or Creutzfeldt-Jakobs diseases. Despite being variable in size, amino acid composition and structure, all amyloidogenic sequences form cytotoxic aggregates [BU15]. The hallmark trait of amyloids is the presence of characteristic short sequences of amino acids, called hot–spots, amyloids can create zipper-like β-structures [F12]. Although studies investigating properties of amyloidogenic sequences were already conducted, the newly established AmyLoad data base facilities large-scale analysis of amyloids [WK15].

Among commonly acclaimed methods of predicting amyloids, only FISH Amyloid [GK14] focuses more deeply 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. Although n-grams constitute a powerful tool for biological sequence analysis, they suffer greatly from a curse of dimensionality. The number of possible n-grams is equal to n^u, where u is the length of the alphabet (4 in case of nucleic acids and 20 in case of proteins). To deal with the abovementioned problem, we we created *biogram* software [BSL15]. Aside from essential functionalities, like efficient data storage, we also implemented a feature selection method. QuiPT (Quick Permutation Test) uses several filtering criteria such as information gained to choose significant features. To speed up the computation and precisely estimate small p-values, QuiPT performs an exact test instead of a large number of permutations.

To reduce the dimension of the problem even more, we group amino acids into bigger groups based on their physicochemical properties important in the aggregation of amyloids. Since it is still unclear which properties are exactly the most important for amylogenicity, we compare classifiers trained on different amino acid aggregations. The grid search not only yields the best aggregation method, but also points which physicochemical properties are important for amylogenicity.

The n-gram model, trained on the data from AmyLoad database, is validated through simple yet accurate amyloid prediction framework using random forests. The preliminary analysis of the amyloidogenic sequences yield not only new insight on the structure of the hot spots, but facilitated prediction of amyloids. The mean AUC of the classifier committee in 5-fold crossvalidation was 0.89.

The predictor of amylogenicity, called AmyloGram, is accessible as a web-server (ADRESS).

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.

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