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. Due to the presence of characteristic short sequences of amino acids, called hot-spots, amyloids can create harmful zipper-like -structures [BU15]. Although studies investigating properties of amyloidogenic sequences were already conducted, the newly established AmyLoad data base facilities large-scale analysis of amyloids [WK15].

To study the data, we use a set of n-gram based methods. N-grams (k-mers) are vectors of n characters derived from input sequences. Originally developed for natural language processing, they are also widely used in genomics, transcriptomics and proteomics. We expect that n-gram analysis, as an addition to commonly acclaimed methods as FISH Amyloid [GK14], will shed more light on the putative motifs of hot spots.

Although n-grams constitute a powerful tool for biological sequence analysis, they suffer greatly from the dimensionality curse. To deal with the abovementioned problem, we we created biogram software [BSL]. 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 allow precise estimation of small p-values, QuiPT performs an exact test instead of a large number of permutations.

Moreover, we aggregate amino acids into bigger groups based on their physicochemical properties important in the aggregation of amyloids. It not only reduces dimensionality of the problem, but preserves relationships between residues. Since it is still unclear which properties are exactly the most important for amylogenicity, we compare classifiers trained on different amino acid aggregations. 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.

References

[BSL] Michal Burdukiewicz, Piotr Sobczyk, and Chris Lauber.

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

[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.