How proteins forming amyloids can be recognized computationally

Malgorzata Kotulska*

Institute of Biomedical Engineering and Instrumentation, Wroclaw Institute of Technology, 50-370 Wroclaw, Poland

*presenting author, e-mail: malgorzata.kotulska@pwr.wroc.pl

Proteins capable of forming fibrils are known as amyloids. Their intramolecular contact sites pattern change in such a manner that they form characteristic zipper aggregates which deprive the protein of its physiological and functional structure. A number of amyloidogenic diseases, which are due to misfolding of a protein into an amyloid fibril, is constantly increasing and include Alzheimer disease (amyloid-β, tau), Parkinson disease (α-synuclein), type 2 diabetes (amylin), Creutzfeldt-Jakob disease (prion protein), Huntington disease (huntington), amyotrophic lateral sclerosis (SOD1), etc. They affect a growing number of people, especially in well developed countries. Recognition of factors responsible for protein misfolding can contribute to better understanding of its mechanisms and potential drug design. Recent studies indicate that short segments of aminoacids, which are called hot spots, can underly amyloidogenic properties of a protein. Those fragments are harmless only when they are burried inside a protein. The amyloidogenic fragments responsible for amyloidogenicity of the whole protein are believed to be 4-10 residues long and it is often assumed that 6-residue fragments of amyloidogenic properties are sufficient “hot spots”. Recognition of amyloidogenic fragments can be obtained by computational approach. In this talk we discuss if classical machine learning methods can be applied for recognition of such hexapeptides and how much we need to know about the training data. We also show an original machine learning method for classification of biological sequences (e.g. sequences of aminoacids), based on discovering a segment with a discriminative pattern of correlations between sequence elements. The pattern is based on location of correlated couples of elements in the window. The algorithm first recognizes the most relevant training segment in each positive training instance. Classification is based on maximal differences between correlation matrices of the relevant segments in positive training sequences and negative training segments. The method efficiency is tested on amyloid proteins.