Abstract

Amyloidogenic regions in polypeptide chains are associated with a number of diseases. Experimental evidence is compelling in favor of the hypothesis that small segments of proteins are responsible for its amyloidogenic behavior. Thus, identifying these short peptides is critical for understanding diseases associated with protein misfolding and developing sequence-targeted anti-aggregation drugs. The in silico approaches using phenomenological models based on bio-physio-chemical properties of amino acids suffer from “curse of dimensionality”. Therefore, before adopting standard classification algorithms to predict such fibril motifs, the “curse of dimensionality” needs to be solved. The present study evaluates the performance of feature selection algorithms namely filter, wrapper and embedded models in conjunction with Support Vector Machine classifier. We also propose a novel integrated feature selection strategy based on Genetic Algorithm and Support Vector Machine to get an optimal
number of features in predicting the amyloid fibril-forming short stretches of peptides. In addition, we investigated the performances of feature selection models that resulted in new and complementary set of properties and concludes that the proposed integrated dimensionality reduction technique outperforms all other methods and achieves the highest sensitivity and specificity of 86% and 82% respectively.

Reference

- http://www.ebi.ac.uk/uniprot/database/download.html
- Jiawei Han, Micheline Kamber. 2008. Data Mining – Concepts and Techniques, Elsevier, II Edition.
- Andries P. Engelbrecht. 2007. Computational Intelligence. John Wiley & Sons Ltd.
An Evaluation of Feature Selection Approaches in Finding Amyloidogenic Regions in Protein Sequences


Kimon K Frousios, Vassiliki A Iconomidou, Carolina-Maria Karletidi, Stavros J Hamodrakas. 2009. Amyloidogenic determinants are usually not buried. BMC Structural Biology. 9:44.


**Index Terms**

Computer Science  
Bioinformatics

**Key words**

Amyloid fibril  
physicochemical properties

Genetic Algorithm

Support Vector Machine