AmylPepPred: Amyloidogenic Peptide Prediction tool

Smitha Sunil Kumaran Nair[1]*, NV Subba Reddy[2] & KS Hareesha[1]

[1]Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal University, Karnataka, India; [2]Mody Institute of Technology and Science University, Rajasthan, India; Smitha Sunil Kumaran Nair – Email: smitha.sunil@manipal.edu; *Corresponding author

Abstract:
We present an efficient computational architecture designed using supervised machine learning model to predict amyloid fibril forming protein segments, named AmylPepPred. The proposed prediction model is based on bio-physio-chemical properties of primary sequences and auto-correlation function of their amino acid indices. AmylPepPred provides a user friendly web interface for the researchers to easily observe the fibril forming and non-fibril forming hexmers in a given protein sequence. We expect that this stratagem will be highly encouraging in discovering fibril forming regions in proteins thereby benefit in finding therapeutic agents that specifically aim these sequences for the inhibition and cure of amyloid illnesses.

Availability: AmylPepPred is available freely for academic use at www.zoommicro.in/amylpeppred

Keywords: Amyloid fibrils, Bio-physio-chemical properties, Auto-correlation function, Support Vector Machine, AmylPepPred

Background:

Amyloid fibril forming proteins are found to be related to amyloid illnesses. Recent experiments suggest that it is not the whole protein; rather short fragments are responsible for amyloidosis [1]. The major limitations of wet lab experimental methods are the time frame involved, high cost and effort. Therefore, a viable solution is through computational approaches. There are web tools available online such as AGGRESCAN [2], AMYLPRED [3], FOLDAMYLOID [4] and so on, but they have varied limitations in maintaining a balance between true positive rates and false positive rates as evaluated [5-7]. AmylPepPred thus provides an open access platform that enables easy and comprehensive retrieval of fibril forming short stretches that compensates the gap in existing amyloid fibril prediction tools by maintaining equilibrium between sensitivity and specificity. This prediction model is a practical implementation of the computational architecture depicted in figure 1 that purely follows a sequence-based design strategy.

Methodology:

The training dataset (Amylpreddataset) has been compiled using experimentally proved proteins related to amyloidosis and proteins with no experimentally determined amyloidogenic regions as described in [6, 7]. The length of wet lab proven positive regions of proteins varies. In fact, the long positive protein segments are broken up into smaller fragments comprising of six amino acids to make the data uniform. Among the 559 properties identified, we extracted a new and complementary set of 40 physicochemical and biochemical properties through Memetic Algorithm, an evolutionary Support Vector Machine (SVM) feature selection approach, besides their auto-correlation function of 5 best pre-selected features in AAIndex database [8] with accession nos. VINM940104, ENGD860101, PRAM900101, KUHL950101, JANJ790101 through SVM within a residue in forming the feature vector to train the SVM model. The overall methodology is illustrated in (Figure 1). The programs are written in C#

BIOINFORMATION

using .NET framework 4.0. The software can be accessed by users from a standard web browser on Windows xp/7 operating system.
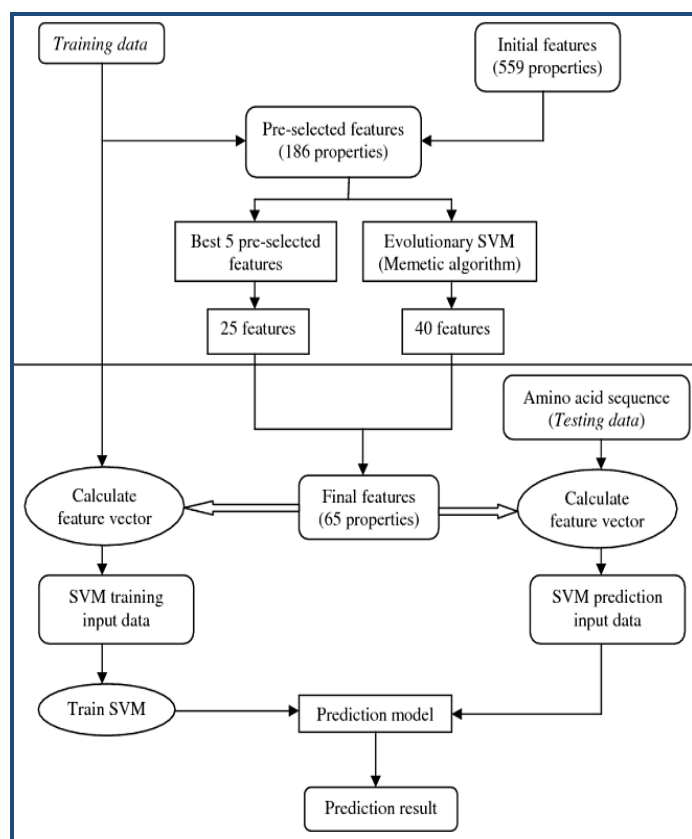
Figure 1: Flowchart illustrating the computational architecture of AmylPepPred

Software input/output:
Once all the related files are downloaded in the same directory, double click the application named, Hexpepfinder. Choose Finder from the menu in the Main window. The user can now browse the input text file containing protein sequence in FASTA format and an output text file. Click Run Finder. The program separates the header and sequence and checks if the input is valid or not. Wait for a pop-up window. To view the output, choose Output file viewer from the menu. By selecting appropriate radio buttons, user can view the fibril forming, non-fibril forming hexmer sequences or both along with positions.

Conclusion:
The study of protein aggregation is crucial to develop rational therapeutic stratagems against amyloid diseases. An encouraging tactic to spot such deposits is through computational prediction models. Nevertheless, these models cannot substitute the wet lab experiments; they might assist in recognizing the regions of concern for further molecular research. AmylPepPred provides a user-friendly interface, a convenient menu driven search option, allowing efficient discrimination of fibril forming and non- fibril forming short protein sequences.

Acknowledgement:

References:
[1] Tian J *et al. BMC Bioinformatics.* 2009 **10**: S45 [PMID: 19208147]
[2] Sole OC *et al. BMC Bioinformatics.* 2007 **8**: 65 [PMID: 17324296]
[3] Frousios KK *et al. BMC Struct Biol.* 2009 **9**: 44 [PMID: 19589171]
[4] Garbuzynskiy SO *et al. Bioinformatics.* 2010 **26:** 3 [PMID: 20019059]
[5] Nair SS *et al. BMC Bioinformatics.* 2011 **12**: S21 [PMID: 22373069]
[6] Nair SS *et al. Bioinformation.* 2012 **8**: 70 [PMID: 22359438]
[7] Nair SS *et al. Protein Pept Lett.* 2012 **19**: 917 [PMID: 22486618]
[8] Kawashima S *et al. Nucleic Acid Res.* 2000 **28**: 374 [PMID: 10592278]