COMPUTATIONAL PREDICTION OF AMYLOIDOGENIC REGIONS IN PROTEINS: A MACHINE LEARNING APPROACH

Smitha Sunil Kumaran Nair[*]

Manipal Institute of Technology, Manipal University, Karnataka, India
[*]*Email: smitha_sunilnair@yahoo.com*

N. V. Subba Reddy

Mody Institute of Technology and Science University, Rajasthan, India
Email: dr_nvsreddy@rediffmail.com

Hareesha K. S

Manipal Institute of Technology, Manipal University, Karnataka, India
Email: hareesh.ks@manipal.edu

Amyloidogenic regions in polypeptide chains are associated with a number of pathologies including neurodegenerative diseases. Recent studies have shown that small regions of proteins are responsible for its amyloidogenic behavior. Therefore, identifying these short peptides is critical for understanding diseases associated with protein aggregation. Owing to the limitations of molecular techniques for the identification of fibril forming targets, it became apparent that clever computational techniques might enable their discovery *in silico*. We propose a machine learning based method to predict the amyloid fibril-forming short stretches of peptides using Support Vector Machine. The features of this method are based on the physicochemical properties of amino acids. To get an optimal number of properties, a feature selection approach based on Genetic Algorithm is performed. The presented algorithm achieved a balanced prediction performance in terms of true positive and false positive rates in predicting a peptide status: amyloidogenic or non-amyloidogenic, which is not reflected in the existing methods.

1. INTRODUCTION

Amyloid fibril formation is widely observed in human diseases including common neurodegenerative pathologies such as, Alzheimer’s disease, Parkinson’s disease, and Huntington’s disease. In these diseases, proteins with unrelated sequences aggregate to form highly characteristic amyloid fibrils. Amyloid fibril formation occurs as a consequence of increase in β strands in amyloidogenic proteins. There is currently no effective treatment against these progressive disorders, most of which affect the brain in a devastating way. Therefore, it is of fundamental medical interest to understand the mechanisms of fibrillogenesis with the ultimate goal of determining the mature fibrils [1].

Recent studies prove that short, continuous and specific amino acid segments act as the major cause of amyloid fibril formation [2]. Therefore, understanding the mechanism of amyloid formation would lead to effective treatments for amyloid illnesses [4].

As reviewed [1], there are many computational approaches used to investigate the regions most prone to form fibrils that result in protein aggregation. A review of existing computational methods for predicting protein aggregates is previously published [19].

Some studies have implied that assembly into amyloid-like fibrils is an inherent property of polypeptides, irrespective of their sequence. However, it is obvious that some sequences are much more amyloidogenic than others. Moreover, some short peptides possess the same amyloid properties as full length proteins, and some very short specific stretches have been considered to be the regions responsible for aggregation, as they can change the amyloidogenic propensities of polypeptides by facilitating or inhibiting fibril formation. These data suggest that peptide sequence can influence amyloid fibril formation, and has inspired the recent development of a number of algorithms and models that predict the amyloidogenic

[*] Corresponding author.

or aggregation propensities of polypeptides or proteins [6].

In this study, our goal is to predict amyloidogenic regions of proteins. The proposed computational method is based on amino acid sequences. We use systematically selected physicochemical properties of amino acids to represent protein sequence features. Genetic Algorithm (GA) is utilized to reduce the dimension of properties. Finally, the Support Vector Machine (SVM) [7] is adopted to classify feature vectors as fibril forming and non-fibril forming peptides.

2. METHOD

2.1. Data set construction

A dataset of six-residue peptides including positive and negative examples of fibril formation is collected from datasets namely Hexpepset [4], AmylHex and AmylFrag [8]. We term this new data set AmylHexpepset and use it to quantify the performance of our method. The Hexpepset dataset consists of 2452 hexpeptides (1226 positive samples and 1226 negative samples). A set of 158 hexmers of which 67 have been shown to form fibrils and 91 have yielded negative results in fibril-forming assays constitute AmylHex. AmylFrag includes a set of 45 amyloidogenic fragments of proteins identified by various researchers. Finally, the AmylHexpepset dataset for training contains 1213 positive samples and 1226 negative samples after removing the discrepancy among the samples in [4, 8] and the redundant samples from the source datasets.

2.2. Feature extraction

As SVM requires each data instance to be represented as a vector of real numbers [7], the numerical values of physicochemical properties of amino acids are used to form the feature vector. These properties are extracted from Amino Acid index database in DBGet (Japan) (AAindex Version 9) [10] and ProtScale in Swiss Expasy [29]. Of the 544 indices in [10], only 216 available in APDbase [30] and 30 in [29] were taken into account for the design. Finally 246 indices are evaluated for potential use.

2.3. Feature selection

Feature selection is a major challenge due to the prevalence of high dimensional data with some irrelevant or redundant features. One of the most fundamental problems in bioinformatics, and machine learning is how to select a small and relevant subset of features. As reviewed [14], there are different feature selection techniques including software packages exist for obtaining minimal feature sets in the field of bioinformatics domain.

The proposed approach of feature selection is a wrapper method based on GA with population size of 10 and predetermined number of 100 generations, wrapped around the classifier, SVM to search for the significant minimal set of features that would improve the overall prediction performance. To achieve a significantly better performance in terms of prediction accuracy, a Perl script is programmed to extract an optimal feature set of 41 properties from 246 properties. All results of GA are obtained using LIBSVM [17] with a 10-fold cross validation on AmylHexpepset.

2.4. Building a model on training data

The SVM classifier is trained with the Kernel RBF. Therefore, all the positive and negative hexpeptides from the training set are implicitly mapped from the input space to a feature space determined by the RBF kernel. In this feature space, an optimal hyperplane is learned by the SVM. In this regard, a suitable setting of the SVM parameter C which is used to control the tradeoff between training error and margin, and the RBF kernel parameter γ that controls the width of the kernel, are determined by a grid search with a 10-fold cross validation on the training dataset. The best set of parameters obtained for the selected feature set are obtained as C=4 and γ=0.25 to achieve best accuracy.

3. PREDICTION ASSESSMENT

In a binary classification, given a classifier and an instance, there are four possible outcomes [18]. When a positive instance is classified correctly as positive, it is counted as a true positive (TP); however if it is classified wrongly as negative, it is counted as a false negative (FN). If the instance is negative and has been classified correctly, it is counted as a true negative (TN), otherwise it is counted as a false positive (FP).

Sensitivity is measured as (TP / (TP + FN)), specificity as (TN / (TN + FP)), Classification accuracy (ACC) as (TP + TN) / (TP + TN + FP + FN) and Matthews Correlation Coefficient (MCC) as (TP * TN – FP * FN) / √ (TN + FN) * (TN+FP) * (TP + FN) * (TP + FP).

4. RESULTS

The model presented in this paper was motivated by the computational challenging task of predicting fibril forming motifs in polypeptide sequences and has been tested on an independent dataset. The algorithm has shown a good balance between sensitivity and specificity in predicting a peptide status compared to existing methods. The presented SVM based model classified the hexpeptides with an overall classification accuracy of .68, Matthews Correlation Coefficient of 0.12, specificity of .71and sensitivity of .51.

5. CONCLUSION

In our present study, kernel-based discriminative method, SVM that uses vector representations of sequences derived from 41 selected sequence properties is used for determining the amyloidogenic stretches in proteins solely from its primary sequence. The proposed method achieves an acceptable result and maintains equilibrium between true positive and false positive rates, when tested on the independent dataset.

The present method is a complement to experimental analysis that may find utility in many medically relevant applications, such as the engineering of protein sequences and the discovery of therapeutic agents that specifically target these sequences for the prevention and treatment of amyloid diseases.

References

1. Amedeo Caflisch: Computational models for the prediction of polypeptide aggregation propensity. Current Opinion in Chemical Biology, *ScienceDirect*, 2007, pp. 437-444.
2. Natalia Sánchez de Groot, Irantzu Pallarés, Francesc X Avilés, Josep Vendrell, and Salvador Ventura: Prediction of "hot spots" of aggregation in disease-linked polypeptides. *BMC Structural Biology* 2005, pp. 5-18.
3. Amol P. Pawar, Kateri F. Dubay, Jesus Zurdo, Fabrizio Chiti, Michele Vendruscolo and Christopher M. Dobson: Prediction of “Aggregation-prone” and “Aggregation-susceptible” Regions in Proteins Associated with Neurodegenerative Diseases. *J. Mol. Bio* 2005,350, pp. 379-392.
4. Jian Tian, Ningfeng Wu, Jun Guo and Yunliu Fan: Prediction of amyloid fibril-forming segments based on a support vector machine. *BMC Bioinformatics* 2009, 10 (Suppl 1): S45.
5. Manuela Lopez de la Paz and Luis Serrano: Sequence determinants of amyloid fibril formation. *PNAS* 2004, Vol. 101, No. 1, pp. 87-92.
6. Zhuqing Zhang, Hao Chen and Luhua La: Identification of amyloid fibril-forming segments based on structure and residue-based statistical potential. *Structural Bioinformatics* 2007, Vol. 23 no. 17, pp. 2218–2225.
7. Christopher J. C. Burges: A tutorial on Support Vector Machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2), 1998, pp. 955-974.
8. Michael J. Thompson, Stuart A. Sievers, John Karanicolas, Magdalena I. Ivanova, David Baker: The 3D profile method for identifying fibril-forming segments of proteins. *PNAS* 2006, Vol. 103, No. 11, pp. 4074–4078.
9. http://www.ncbi.nlm.nih.gov/protein/
10. Kawashima S, Kanehisa M: AAindex: amino acid index database. *Nucleic Acids Res* 2008, 28(1): 374.
11. Jiawei Han, Micheline Kamber: Data Mining – Concepts and Techniques, Elsevier 2008, II Edition.
12. Laskko T A, Bhagwat J G, Zou K H, Ohno Machado L: The use of receiver operating characteristic curves in biomedical informatics. *J Biomed Inform* 2005; 38: 404-415.
13. Kudo M, Sklansky J: Comparison of algorithms that select features for pattern recognition. *Pattern Recognition* 2000, 33(1): 25-41.
14. Ferri F J, Pudil P, Hatef M, Kittler J: Comparative study of techniques for large-scale feature selection. Pattern Recognition in Practice IV, *Elsevier*, 1994, pp.403-413.
15. Yvan Saeys, Inaki Inza, Pedro Larran: A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007, Vol. 23 no. 19, pp. 2507–2517.

16. Andries P. Engelbrecht: Computational Intelligence. *John Wiley & Sons Ltd. Publishers*, II Ed, 2007.
17. http://www.csie.ntu.edu.tw/~cjlin/
18. Pierre Baldi, Soren Brunak, Yves Chauvin, Claus A F Anderson, Henrick Nielson: Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 2000, Vol 16, No. 5.
19. Smitha Sunil Kumaran Nair, N. V. Subba Reddy, Hareesha K. S: Computational models for the prediction of amyloid fibril forming protein segments. *Proc. Int'l Conference on Bioinformatics and Systems Biology* 2010, Annamalai University, Vol. 1, pp.152-157.
20. Kimon K Frousios, Vassiliki A Iconomidou, Carolina-Maria Karletidi, Stavros J Hamodrakas: Amyloidogenic deteminants are usually not buried. *BMC Structural Biology* 2009, 9:44.
21. Oxana V. Galzitskaya, sergiy O. Garbuzynskiy, Michail Yurievich Lobanov: Prediction of Amyloidigenic and Disordered Regions in Protein Chains. *PLoS Computational Biology* 2006, Volume 2, Issue 12, e177.
22. Magdalena I. Ivanova, Michael J. Thompson, and David Eisenberg: A systematic screen of β2-microglobulin and insulin for amyloid-like segments. *PNAS* 2006, Vol. 103, No. 11, pp. 4079–4082.
23. Ana-Maria Fernandez-Escamilla, Frederic Rousseau, Joost Schymkowitz & Luis Serrano: Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature Biotechnology* 2004, Vol. 22, No. 10, pp. 1302-1306.
24. http://www.mobioinfor.cn/pafig
25. http://antares.ru/fold-amyloid/
26. Oscar Conchillo-Sole, Natalia S de Groot, Francesc X Aviles, Josep Vendrell, Xavier Daura and Salvador Ventura: AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics* 2010, 8:65.
27. Susan Idicula-Thomas and Petety V Balaji: Understanding the relationship between the primary structure of proteins and their amyloidogenic propensity: clues from inclusion body formation. *Journal of Protein Engineering, Design & Selection* 2005, Vol. 18, No. 4, pp. 175-180.
28. Mathura & Kolippakkam: APDbase: Amino acid Physicochemical properties Database. *Bioinformation* 2005, 1(1): 2-4.
29. http://www.expasy.org/tools/protscale.html
30. http://www.rfdn.org/bioinfo/APDbase.php