Prediction of Amyloid Fibrillar Aggregates of Polypeptide Sequences: A Soft Computing Approach

Smitha Sunil Kumaran Nair, *Member, IAENG*,
N. V. Subba Reddy, Hareesha K. S, Sunil Kumaran Nair S

***Abstract*—The deposition of amyloid fibrillar aggregates in human brain results in amyloid illnesses. As these aggregates may spread like virus, it is of primary importance to spot such motif regions in protein sequences. Limitations of molecular techniques in identifying them offer sophisticated computational methods for their efficient retrieval. In this paper we tried to enhance the prediction performance of computational approaches by the union of machine learning algorithms: an approach from a soft computing perspective. A filter based dimensionality reduction algorithm has been utilized on the extracted features to obtain a minimal feature subset for Decision tree classification. The filter approach is a multivariate statistical analysis based on the mutual information which is a mixed measure of maximum Relevance and Minimum Redundancy of features. We performed stratified 10-fold cross-validation test to objectively evaluate the accuracy of the predictor.**

***Index Terms*—Amyloid fibrillar aggregates, Atomic composition, physio-chemical properties, maximum Relevance Minimum Redundancy, Decision tree**

I. Introduction

IT has been established that amyloid fibrillar aggregates of polypeptide sequences are associated with amyloid illnesses such as Alzheimer's disease [1]. Recognizing the factors to slow down or completely prevent such devastations, the researchers have intensified their efforts to investigate ways to delay, cure, or prevent the onset and progression of diseases. The experimental validation of peptide segments prone to form fibrils is tedious; hence it is imperative that machine learning approaches would be significant before wet lab experiments are carried out even though the computational methodologies cannot fully substitute molecular techniques [2].

Several computational techniques based on various parameters for predicting amyloid fibrillar aggregates exist namely FoldAmyloid [3], Aggrescan [4], Amylpred2 [5], Zipperdb [6] and Pafig [7]. We carried out an extensive assessment on few tools mentioned above [8] and found that the prediction accuracies are still far from satisfied.

In this paper, we tried to present a soft computing approach to predict amyloid fibrils in protein sequences. The data samples are encoded by 5 values of atomic composition and 50 physio-chemical properties selected by the feature optimization method namely maximum Relevance and Minimum Redundancy (mRMR) to train a Decision tree (DT) classifier. Section II covers the materials and methods utilized for the present study. Results and their analysis are discussed in section III followed by concluding remarks.

Smitha Sunil Kumaran Nair is with the Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal University, Karnataka, India (e-mail: smitha.sunil@manipal.edu).

N. V. Subba Reddy is with Mody Institute of Technology and Science University, Rajasthan, India (e-mail: dr_nvsreddy@rediffmail.com).

Hareesha K. S is with the Department of Computer Applications, Manipal Institute of Technology, Manipal University, Karnataka, India (e-mail: hareesh.ks@manipal.edu).

Sunil Kumaran Nair S is with the Department of Management Studies, Annamalai University, Tamilnadu, India (email: zunilnair@yahoo.com).

II. Materials and Methods

A. Sequence data preparation

The performance of a soft computing algorithm improves by training the model with appropriate datasets. As suggested by Conchillo-Sole et al. [4], the success of computational approaches in predicting aggregation-prone regions allow proposing that aggregation propensity in polypeptide chains is dictated by amino acid sequences. Moreover, various independent investigations point out that the capability of a protein to be amyloidogenic is concentrated in certain regions and, more precisely, in small sequence fragments [9] termed hotspot/motif which shows a key role in the transformation of proteins from their soluble state into fibrillar, β-structured aggregates. Thompson et al. [10], claimed that a combination of six amino acids is sufficient to form amyloid-like fibrils. Therefore, a dataset of hexapeptides have been utilized provided by Tian et al. [7]. This dataset contains 2452 samples with an equal distribution of positive and negative hexamers. Fig. 1. shows samples of fibril forming hexamers prepared from a protein sequence.

Fig. 1. Protein sequence and fibril forming hexamers

B. Feature extraction and formulation

In the context of supervised machine learning, the general goal of a machine learning task is to map input data samples into some output labels. To achieve this, the soft computing model should be trained by samples described by features.

ISBN: 978-988-19252-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

In this work, we have considered two features: atomic composition and physio-chemical properties of amino acids.

Atomic composition (AC) is a feature considered for feature vector representation. This refers to Carbon, Hydrogen, Nitrogen, Oxygen and Sulphur atoms that make up an amino acid sequence. This corresponds to a 5-dimensional feature vector formation. As the count of constituent atoms in each hexamer varies from one another, this feature is hypothesised to be a good choice as it helps in differentiating samples. Mathematically, this can be formulated as follows.

Let $A = \{S, O, N, H, C\}$ be the atoms. AC of a hexamer sequence h_i is calculated as $A(h_i) = \{a_1, a_2, a_3, a_4, a_5\}$ where $a_i \in A,$ that refers to the count of each atom type in a hexamer h_i.

As suggested by Pastor et al. [11], one of the vital factors that influence amyloid deposition owes to the determination of the physio-chemical properties of amino acids. We have extracted 531 properties associated with each of the 20 amino acids available in AAIndex database [12].

C. Feature selection

Any learning agent must learn from experience to discriminate between the relevant and irrelevant fragments of its experience: a ubiquitous problem [13]. Feature optimization algorithms are faced with the problem of selecting subset of features upon which to focus its attention, while ignoring the rest. To achieve a considerably better performance in terms of prediction ability, it is a prerequisite to select relevant features so as to discriminate well among classes.

As the total number of extracted features is 536, there is a requirement to use an effective feature reduction technique to obtain a minimal feature subset. In this paper, we utilized a multivariate filter method namely maximum Relevance Minimum Redundancy.

The purpose of feature selection is to find a feature subset S with m features $\{x_i\}$. This subset should have the largest dependency on the target class c. Mathematically, this scheme, namely Max-Dependency, takes the form:

$$\max D(S,c); D = I(\{x_i, i = 1,2,--,m\}; c) \quad (1)$$

As suggested by Peng et al. [14], Max-Dependency is difficult to implement, therefore an alternate approach is to select features based on Max-Relevance. In this method, search features that satisfy equation (2) which approximates $D(S,c)$ in equation (1) with the average values of all the mutual information values between a single feature x_i and target class c.

$$\max D(S,c); D = \frac{1}{|S|}\sum_{x_i \in S} I(x_i; c) \quad (2)$$

However, there is a chance of Max-Relevant features to contain redundancy. Hence, an approach of Min-Redundancy can be incorporated to select mutually exclusive features according to equation (3).

$$\min R(S); R = \frac{1}{|S|^2}\sum_{x_i, x_j \in S} I(x_i, x_j) \quad (3)$$

The above criteria are combined to form minimal Redundancy Maximum Relevance. An operator defined by

$$\psi(D,R); \psi = D - R \quad (4)$$

ψ is used to combine D and R. Equation (4) is utilized to optimize D and R simultaneously.

As a result of applying this technique, the first 50 features were selected based on highest scores. Finally, the feature vector was constructed with 55 values (5 atomic composition and 50 physio-chemical properties).

D. Soft computing model

DT classifier is used to train and test the data samples encoded by a feature vector. This classifier is based on the results of a series of tests carried out on the attributes of a sample. It works by posing a series of questions about the attributes associated with unknowns; each question is contained in a node, and each node has child nodes for each possible answer to its question. It eventually terminates in leaves, which correspond to a classification. The state-of-art implementation of Alternating Decision Tree in Weka data mining package [15] is utilized that supports two-level classification. The number of boosting iterations was manually tuned to 10 so as to suit the dataset and the desired complexity/accuracy tradeoff. More boosting iterations may result in larger (potentially more accurate) trees, but possibly make learning slower [16]. Each iteration adds 3 nodes (1 split + 2 predictions) to the tree unless merging occurs. All other parameters were set to default values.

III. RESULTS AND DISCUSSION

The soft computing approach that has been adopted to predict amyloid fibrillar aggregates of polypeptide sequences is depicted in fig. 2 and fig. 3. Fig. 2. consists of positive and negative samples in the form of training dataset encoded by the feature values to train a soft computing model. This results in a knowledge base. Once the knowledge base is obtained, the testing sequences encoded with feature values are tested to predict if a sequence belongs to a category of amyloidogenic or not.

Accurate *in silico* prediction methods of amyloidogenic peptide regions rely on the cooperation between data samples, informative features and classifier design. In this study, we have used a dataset related to amyloid aggregates detected by previous proteomic studies. We exploited heterogeneous features based on physio-chemical properties and atomic composition. mRMR [17] has been used to reduce the dimensionality of the feature vector. Finally a DT classifier is trained to obtain the knowledge base.

ISBN: 978-988-19252-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

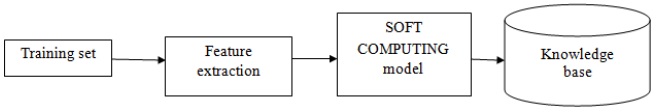

Fig. 2. Training process of hexamer samples

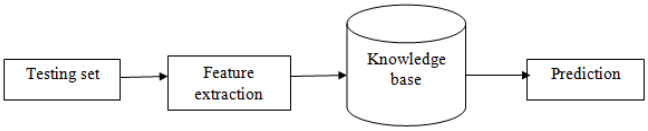

Fig. 3. Prediction of test samples

Statistical metrics namely Sensitivity (S_n) and Specificity (S_p) have been used as accuracy indices to evaluate the performance of the predictor. In biomedical statistics, S_n is the probability of correctly predicting a positive example calculated as $S_n = \frac{TP}{TP + FN}$ and S_p is the chance of correctly predicting a negative example computed as $S_p = \frac{TN}{FP + TN}$. In a binary classification, given a classifier and an instance, there are four possible outcomes. 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) [18].

Stratified 10-fold cross validation is the evaluation protocol used to compute the true positive and false positive rates as it is generally recommended for estimating accuracy due to its relatively low bias and variance [19]. Scores of .75 and .76 have been obtained for S_p and S_n respectively for the proposed approach.

It has been found that the prediction model trained with the individual features considered under study resulted in a considerable reduction in detecting true negatives and false positives. However, acceptable accuracy could be obtained by assimilating diverse feature elements in predicting amyloid aggregates. Even though the results of proposed model match favorably with other methods, it needs to be enhanced further. Improvement in prediction results may be possible by identifying more relevant features, and/or by incorporating more relevant training data.

IV. CONCLUSION

Prediction of amyloid-like fibril aggregates in polypeptide sequences is important in understanding the underlying cause of amyloid illnesses. This will aid in the discovery of sequence-targeted anti-aggregation pharmaceuticals. Owing to the constraints of molecular wet lab experiments for their identification, it is highly desirable to develop computational architectures to provide better and affordable *de-novo* predictions. In this paper, we propose a soft computing approach to predict several previously "hidden" protein segments that tend to aggregate resulting in amyloid diseases. Attaining the knowledge of such regions is important to comprehend their abnormal assemblage and deposition, which would perhaps lead to the rational design of therapeutic targets for their inhibition or disaggregation.

REFERENCES

[1] L. Goldschmidt, P. K. Teng, R. Riek, D. Eisenberg, "Identifying the amylome, proteins capable of forming amyloid-like fibrils," PNAS 107, 2010, No.8, pp. 3487-3492.

[2] S.S.K. Nair, N.V. S. Reddy, K. S. Hareesha, "Exploiting heterogeneous features to improve in silico prediction of peptide status – amyloidogenic or nonamyloidogenic," BMC Bioinformatics, 12(Suppl 13), 2011, S21.

[3] S. O. Garbuzynskiy, M. Y. Lobanov and O. V. Galzitskaya, "FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence," Structural Bioinformatics, vol. 26, 2010, No.3, pp. 326-332.

[4] O. Conchillo-Sole, N. S. de Groot, F. X. Avilés, J. Vendrell, X. Daura, and S. Ventura, " AGGRESCAN: a server for the prediction of "hot spots" of aggregation in polypeptides," BMC Bioinformatics,8, 2007, 65.

[5] A. C. Tsolis, N. C. Papandreou, V. A. Iconomidou, S. J. Hamodrakas, "A Consensus Method for the Prediction of 'Aggregation-Prone' Peptides in Globular Proteins," PLoS ONE, 8(1), 2013, e54175.

[6] http://services.mbi.ucla.edu/zipperdb
This database contains predictions of fibril-forming segments within proteins identified by the 3D Profile Method in [10].

[7] J. Tian, N. Wu, J. Guo and Y. Fan, "Prediction of amyloid fibril-forming segments based on a support vector machine," BMC Bioinformatics, 10(Suppl 1), 2009, S45.

[8] S. S. K. Nair, N. V. S. Reddy, K. S. Hareesha, "Motif mining: an assessment and perspective for amyloid fibril prediction tools," Bioinformation, 8, 2012, (2), pp. 070-074.

[9] A. S. Reddy, M. Chopra, and J. J. de Pablo, "GNNQQNY—Investigation of Early Steps during Amyloid Formation," Biophysical Journal, Vol. 98, 2010, pp. 1038–1045.

[10] M. J. Thompson, S. A. Sievers, J. Karanicolas, M. I. Ivanova, D. Baker, "The 3D profile method for identifying fibril-forming segments of proteins," PNAS, Vol. 103, 2006, No. 11, pp. 4074–4078.

[11] M. T. Pastor, A. Esteras-Chopo and M. L. de la Paz, "Design of model systems for amyloid formation: lessons for prediction and inhibition," Current Opinion in Structural Biology, 15, 2005, pp. 57–63.

[12] S. Kawashima, M. Kanehisa, "AAindex: amino acid index database," Nucleic Acids Res, 28, 2008, (1), pp. 374.

[13] R. Kohavi, G. H. John, "Wrappers for Feature Subset Selection," AIJ special issue on relevance, 1997, pp. 1-43.

[14] H. Peng, F. Long, and C. Ding, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, 2005, No. 8, pp. 1226-1238.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, "The WEKA Data Mining Software: An Update," SIGKDD Explorations, 11, 2009,1.

[16] S. S. K. Nair, N. V. S. Reddy, K. S. Hareesha, "Machine learning study of various classifiers trained with biophysiochemical properties of amino acids to predict fibril forming motifs in peptides," Protein and Peptide Letters, Vol. 19, 2012, No. 9, pp. 917-923

[17] http://penglab.janelia.org/proj/mRMR/
This helps in selecting the features with minimum-Redundancy-and Maximum-Relevance as described in [14].

[18] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Anderson, H. Nielson, "Assessing the accuracy of prediction algorithms for classification: an overview," Bioinformatics Review, Vol 16, 2000, No. 5, pp. 412-424.

[19] J. Han and M. Kamber, "Data Mining: Concepts and techniques," Morgan Kaufmann Publishers, 2006.

ISBN: 978-988-19252-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)