Assessing the Accuracy of Computational Tools for the Prediction of Amyloid Fibril forming Motifs: An Overview

Smitha Sunil Kumaran Nair
Department of Computer Science and Engineering
Manipal Institute of Technology
Manipal University, Karnataka, India

N. V. Subba Reddy
Mody Institute of Technology and Science University
Rajasthan
India

Hareesha K. S
Department of Computer Science and Engineering
Manipal Institute of Technology
Manipal University, Karnataka, India

ABSTRACT

Identifying amyloidogenic regions in protein sequences is useful in understanding the underlying cause of several human diseases and finding potential therapeutic targets. Given the laborious nature of experimental validation of segments most prone to form fibrils, it was essential that computational approaches be developed that could produce reliable, affordable and testable *in silico* predictions. In this paper, we present and assess some of the recently developed computational tools for predicting amyloid fibril forming motifs that remain as one of the key means used to decipher the role of such regions in disease diagnosis, prognosis and drug discovery.

Keywords

Amyloid fibrils, Computational tools, Prediction accuracy

1. INTRODUCTION

A large number of diseases caused by the aggregation of misfolded proteins results in the formation of amyloid fibrils. Numerous studies have shown that, in addition to proteins involved in amyloid diseases, many proteins not related to any amyloid disease can aggregate into fibrils under destabilizing conditions [1]. However, it is obvious that some sequences are much more amyloidogenic than others [2]. Evidence indicates that short sequence stretches within a protein primary sequence may be responsible for amyloid formation [3].

The focus of this review is on recent approaches to predict the amyloidogenic motifs of polypeptide sequences. Many groups have actively worked on developing tools that integrate several factors driving protein aggregation in order to identify potential amyloidogenic stretches in proteins. Most of the methods were published in 2004-2010 and show a good agreement with experimental results.

2. OVERVIEW OF PREDICTION TOOLS

The challenge of predicting amyloidogenic regions has resulted in a variety of multi-parametric methods that attempt to predict such motif sequences. Each method makes its own hypothesis and implements, which range from quite simplistic to quite complex [7]. Overall, the success of different computational approaches in predicting aggregation-prone regions allow proposing that aggregation propensity in polypeptide chains is ultimately dictated by the sequence [6]. Here we summarize some of the prediction models capable of discriminating between amyloidogenic peptides and non-amyloidogenic peptides that are purely based on primary structure of proetins. Few of these amyloidogenic motif mining tools are made available through online resources as mentioned in the subsequent sections.

2.1 3D Profile method

A computational approach is developed based on the crystal structure of the cross-β spine formed by the peptide NNQQYY, for identifying those segments of amyloidogenic proteins that themselves can form amyloid-like fibrils [5]. Their approach is built on experiments showing that a group of six amino acids are sufficient for forming amyloid fibrils. A sequence of interest is scanned by sliding a window of six residues and maps each peptide onto templates of the crystal structure of the NNQQYY peptide. Each mapping of sequence to template is evaluated energetically with ROSETTADESIGN [4] and the prediction is made by taking the best scoring fit between peptide and template. The putative prediction is accepted as a prediction if its energy is lower than the threshold energy. According to Thompson *et al.*, even though the presented template method shows promise in discriminating between fibrils and non-fibrils especially in tau proteins and myoglobin, this may still be improved as more template structures become known.

2.2 PreAmyl

Zhang et al., use structure and residue-based statistical potential for the identification of amyloid fibril forming segments. A template library is constructed with 2511 structures with a slight perturbation in coordinates of the microcrystal structure of the NNQQYY peptide. Each expected hexpeptide is mapped onto each of the template structures. The residue-based statistical potential (statistical mean force extracted from experimentally solved protein structures) is used to evaluate the interaction energy scores. The lowest energy score obtained from the template structures is then used to assay the fibril forming propensity of this peptide [2]. Examination of proteins related to amyloidosis agrees with experimental data. In fact, the major limitation of pre-amyl lies in that only the microcrystal X-ray structure of NNQQYY was used, which does not show common fibril twists and predictive power may be improved further with more experimental structure models.

2.3 Aggrescan

Aggrescan (O. C. Sole et al., 2007) [6] is a web-based software that can predict aggregation-prone segments in protein sequences. Using an in vivo reporter method to study a "hot spot" in the central hydrophobic core of Aβ, the effect of single point mutations on the aggregation propensities of the peptide within the cell is calculated. The results are used to approximate the in vivo intrinsic aggregation propensities of natural amino acids when located in an aggregation-prone sequence stretch. This information was subsequently used to generate an aggregation profile for any protein sequence under study to detect those regions with high aggregation propensities. Identification of such regions is accessed through the link http://bioinf.uab.es/aggrescan/.

2.4 AmylPred

A publicly available online tool that utilizes five different and independently published methods, to form a consensus prediction of amyloidogenic regions in proteins, using only protein primary structure data is developed (Frousios *et al.,* 2009) [7]. The first method relies on average packing density profiles. The second method used is the consensus secondary structure prediction algorithm SecStr [8] that has been shown to be able to predict amyloidogenic regions as conformational switches, which are identified as regions predicted both as α-helices and β-strands. Locating the amyloidogenic pattern {P}-{PKRHW}-[VLSCWFNQE]-[ILTYWFNE]-[FIY]-{PKRH} [9] is another method used for the consensus prediction. The TANGO algorithm [10] based on the physicochemical principles underlying β-sheet formation, extended by the assumption that the core regions of an aggregate fully buried, is the next method used (version 2.1) that calculates the tendency of peptides to form beta aggregates and aside from the primary sequence. Finally, an algorithm that maps all hexapeptides of a sequence onto the microcrystalline structure of NNQQNY and calculates the resulting conformational energy [2] is used. The tool is available at http://biophysics.biol.uoa.gr/AMYLPRED/ input.html.

2.5 Pafig

The method, named Pafig (Prediction of amyloid fibril-forming segments) based on support vector machines is proposed, to identify the hexpeptides associated with amyloid fibrillar aggregates [11]. The predictive model of Pafig is a phenomenological model, based on 41 physicochemical properties selected by a two-round selection from 531 physicochemical properties in the Amino acid index database (AAindex) [12]. Because short regions of a protein are responsible for its amyloidogenic behavior, Pafig was trained by hexpeptides, which were decomposed by scanning for segments that could form fibrils with a six-residue sliding window. Using a 10-fold cross validation test on Hexpepset dataset, Pafig performs with an overall accuracy of 81%.

The features of Pafig do not contain the structural features of the proteins. Thus, it is possible that some of the structure information is ignored by Pafig, which also exist as the protein aggregation and fibril forming factors. However, the lack of structural information is likely to overcome by the inclusion of different physicochemical properties in the Pafig. Moreover, the sample size of the training dataset of Pafig compared with the number of all hexpeptides is small, which would affect the performance of Pafig. Therefore, collection of more data by combining biological knowledge and related sources and integration of some structure features into Pafig would improve its prediction accuracy rate.

2.6 FoldAmyloid

FoldAmyloid (S. O. Garbuzynskiy et al., 2010) [1] tool is based on using expected characteristics – scales: either expected packing density or the probability of formation of hydrogen bonds. The scales themselves are obtained from the statistics of spatial structures of proteins, and then the scales are used for predictions on amino acid sequences. Initially, the values of the expected packing density and probability of formation of hydrogen bonds for each residue in spatial structures of proteins are obtained. The average values for each of 20 types of amino acid residues are calculated. The obtained average values are then used as the values expected for each residue of a given type in a sequence for which the prediction is made. The FoldAmyloid web server is available at http://antares.protres.ru/fold-amyloid/.

3. DISCUSSION AND CONCLUSION

In the absence of high-throughput experimental techniques to determine the fibril forming regions, it is vital that computational techniques are developed to unravel their effects in protein misfolding and implications for disease diagnosis. Reliable predictions of amyloidogenic regions have a great impact in the development of anti aggregation drugs.

To evaluate the quality of each prediction tools, we compiled experimentally proved proteins related to amyloidosis and proteins with no experimentally determined amyloidogenic regions published in literature [1, 2, 5, 6, 11, 13, 14, 15] in order to construct the dataset. We compared a few tools such as Aggrescan, Amylpred and FoldAmyloid to evaluate the performance of their predictability based on the prepared dataset. Of all the tools examined, Aggrescan achieves the best overall prediction accuracy in terms of sensitivity and specificity. In addition, a significant reduction of sensitivity associated with a gain in specificity is noted in all the tools considered under the present study. Therefore, even though the algorithms for their predictions have improved over time, accurate predictions still remain a challenging task and are still subject of intense investigations.

4. REFERENCES

[1] Sergiy O. Garbuzynskiy, Michail Yu. Lobanov and Oxana V. Galzitskaya, " FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence," Struct Bioinformatics 2010, Vol. 26, No. 3, pp. 326-332.

[2] 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.

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

[4] http://www.rosettacommons.org

[5] 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.

[6] Oscar Conchillo-Sole, Natalia S de Groot, Francesc X Avilés,Josep Vendrell, Xavier Daura, and Salvador Ventura, "AGGRESCAN: a server for the prediction of "hot spots" of aggregation in polypeptides," BMC Bioinformatics 2007, 8:65.

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

[8] Hamodrakas SJ: A protein secondary structure prediction scheme for the IBM PC and compatibles. Comput Appl Biosci 1988, 4:473-477.

[9] López de la Paz M, Serrano L: Sequence determinants of amyloid fibril formation. Proc Natl Acad Sci 2004, 101:87-92.

[10] Fernandez-Escamilla AM, Rousseaux F, Schymkowitz J, Serrano L: Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nature Biotechnology 2004, 22:1302-1306.

[11] Jian Tian, Ningfeng Wu, Jun Guo and Yunliu Fan, "Prediction of amyloid fibril-forming segments based on a support vector machine," *BMC Bioinformatics*, January 2009, **10**(Suppl 1):S45.

[12] Kawashima S, Kanehisa M, "AAindex: amino acid index database," *Nucleic Acids Res* 2000, 28(1):374

[13] 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, 5:18, doi:10.1186/1472-6807-5-18.

[14] 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.

[15] Sukjoon Yoon, William J. Welsh, "Detecting hidden sequence propensity for amyloid fibril formation, "Protein Science 2004, 13: 2149-2160.