BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 8(2)

Motif mining: an assessment and perspective for amyloid fibril 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

Received December 07, 2011; Accepted December20, 2011; Published January 20, 2012

Abstract:
Amyloid fibril forming regions in protein sequences are associated with a number of diseases. Experimental evidences compel in favor of the hypothesis that short motif regions are responsible for its amyloidogenic behavior. Thus, identifying these short peptides is critical in understanding the cause of diseases associated with aggregation of proteins and developing sequence-targeted anti-aggregation drugs. Owing to the constraints of wet lab molecular techniques for the identification of amyloid fibril forming targets, computational methods are implemented to offer better and affordable *in silico* predictions. The present study takes into consideration an assessment and perspective of the recent tools available for predicting a peptide status: amyloidogenic or non-amyloidogenic. To the best of our knowledge, the existing review articles on amyloidogenic prediction tools have not touched upon their effectiveness in terms of true positive rates or false positive rates. In this work, we compare few tools such as Aggrescan, Amylpred and FoldAmyloid to evaluate the performance of their predictability based on the experimentally proved data in terms of specificity, sensitivity, Matthews Correlation Coefficient and Balanced accuracy. As evident from the results, a significant reduction of sensitivity associated with a gain in specificity is noted in all the tools considered under the present study.

Background:
Amyloid fibril forming regions in protein sequences appear to be associated with several illnesses including neurodegenerative diseases and Type II diabetes **[1]**. Experimental proof is compelling in approval of the postulate that continuous and short segments of peptide sequences are responsible for amyloidogenity **[1, 2]**. Hence mining such motifs is important in understanding the underlying cause of amyloid illnesses. The reliable discovery of amyloid promoting fragments in proteins has a great impact on the development of anti-amyloid agents as well. Moreover, methods that identify aggregation-prone motifs have a broad range of biotechnological applications, such as the improvement of the solubility of recombinant proteins for pharmaceutical and industrial purposes, and peptide-based biomaterial engineering **[3]**. Therefore, during the past five years, 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 show a good agreement with wet lab experimental results.

Recent efforts in understanding the physicochemical grounds **[4]** and structural denominators **[5]** of amyloid fibril formation has led to the development of several algorithms, capable of predicting a number of aggregation related parameters of a protein directly from its amino acid sequence. Review articles on computational studies of investigating fibril forming segments do exist such as **[2]**, but are focused on the design of model systems for amyloid formation **[3]**, approaches based on computer simulations of the aggregation process of proteins **[6]** and those emphasizing on phenomenological models that use the physicochemical properties of the side chains and computational techniques based on atomistic descriptions of β–aggregation **[5]**. In fact, these articles have not touched upon the effectiveness of prediction tools in terms of true positive rates or false positive rates. Therefore, the main focus of our present study involves the evaluation of recent computational prediction tools to predict amyloidogenic stretches of polypeptide sequences based on statistical parameters. However, we believe that such a comparative analysis of fibril

forming prediction tools might be useful to carry out further research in this area.

Methodology

The challenge of computationally mining amyloidogenic regions has resulted in a diversity of multi-parametric methods that attempt to predict fibril motifs **[7]**. The problem remains that many methods are not available to be downloaded for inclusion in independent testing on a common dataset. Hence, at the moment, we have taken into account of only those most recent methodologies which provide an online tool solely based on sequences as input to verify the amyloidogenity of a peptide. Few other methods such as 3D profile method **[8]** based on the crystal structure of the cross-β spine formed by the peptide NNQQNY, PreAmyl **[9]** based on structure and residue-based statistical potential, Pafig **[1]** based on supervised learning model trained with 41 physicochemical properties, are not included in this review due to their incompatibility with the present study.

Data retrieval and preparation

The overall success of diverse computational approaches in predicting aggregation-prone regions allows to propose that aggregation propensity in polypeptide chains is ultimately dictated by the primary protein sequences **[10]**. The quality of each prediction tool has been evaluated on two datasets namely Amylpreddataset and AmylFibrilset. Amylpreddataset corresponds to the data available in **[11]**. Frousios *et al*. compiled 18 proteins (Accession Nos.: P01236, P01258, P02647, P02663, P02735, P02766, P02788, P04156, P04279, P05067, P06396, P10636, P10997, P22398, P37840, P61626, P61769 and Q08431) having experimentally proved fibril regions and 5 proteins (Accession Nos.: P00441, P01034, P01308, P01857, P01625) which showed no signal of fibrillogenesis. Besides, we compiled experimentally proved proteins related to amyloidosis and proteins with no experimentally concluded amyloidogenic regions published in literature **[1, 8-16]**, in order to construct a dataset. We term this dataset, AmylFibrilset. AmylFibrilset includes natively globular proteins, natively intrinsically unstructured proteins, amyloidogenic proteins and proteins related to depositional diseases to analyze deeper the general predictive ability of each method. The accession numbers of proteins included in this dataset can be retrieved from our earlier publication **[17]**.

Recent findings in the study of protein aggregation reveal the fact that there exist rather specific continuous small stretches that can nucleate the aggregation process **[10]**. Thompson *et al*. **[8]** claim that a hexmer is sufficient to form amyloid-like fibril motifs. Therefore, the predicted peptides with less than 6 residues were excluded while quantifying each method. The total data in AmylFibrilset collectively amounts to 10,603 six amino acid residues obtained by a six-residue sliding window. The total amount of positive residues is 1176, experimentally found to aggregate and 917 peptides known not to aggregate by experiment. Amylpreddataset contains 512 positive hexmers and 829 negative hexmers selected from 6,761 hexpeptides. The use of positive dataset helps to identify the number of true positives and false negatives that defines the sensitivity of a particular tool. The false positives and true negatives defining the specificity of a tool are obtained using negative dataset.

Analysis

Here in the present study, we summarize, in alphabetical order, three prediction programs capable of discriminating between amyloidogenic peptides and non-amyloidogenic peptides **Table 1 (see supplementary material)**. The positive and negative data that have been experimentally supported and obtained from literature mining as detailed above were given as input to each of these tools and the predicted output was analyzed to calculate the count of true positives, true negatives, false positives and false negatives **Table 2 (see supplementary material)**. Further, to evaluate the performance of prediction tools, we calculated the statistical parameters namely Sensitivity, Specificity, Matthews Correlation Coefficient and Balanced accuracy. The comparative analysis tells how well each tool can predict the fibril motifs in a given sequence.

Aggrescan

Aggrescan **[10]** is 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/ **[18]**. The analysis was performed using the default parameters of the software.

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 **[11]**. The first method relies on average packing density profiles. The second method used is the consensus secondary structure prediction algorithm SecStr **[19]** 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} **[12]** is another method used for the consensus prediction. The TANGO algorithm **[20]** 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 hexpeptides of a sequence onto the microcrystalline structure of NNQQNY and calculates the resulting conformational energy **[9]** is used. The tool is available http://biophysics.biol.uoa.gr/AMYLPRED/input.html**[21]**.The analysis was performed using the default parameters for each employed algorithm.

FoldAmyloid

FoldAmyloid **[22]** algorithm 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/ **[23]**. We chose the value of sliding window size and reliable frame size to be 6 to carry out the analysis.

Results & Discussion:

Multiple statistical measures were used to assess the performance of the tools under study including Sensitivity (S_n), Specificity (Sp), Matthews Correlation Coefficient (MCC) **[24]** and Balanced accuracy (BACC) **[25]** related to the standard Balanced Error Rate (BER), where *BER = 1-BACC* . 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). The MCC used in machine learning is a measure of the quality of binary classifications. It takes into account true and false positives and negatives and is generally regarded as a balanced measure. It returns a value between -1 and +1. A coefficient of +1 represents a perfect prediction, 0 an average random prediction and -1 the worst possible prediction **[24].**

The methods included in our analysis were not assigned any prediction cutoff or threshold. As a result, Receiver operating Characteristic (ROC) curves cannot be completely constructed for these algorithms. Instead, the overall prediction performance of these tools could only be represented by selected points on the ROC plot. Figure 1 shows the scatter plot for true positive rate (sensitivity) versus false positive rate (1-specificity) of each of the prediction tools and the balance between S_n and S_p of these methods are compared. The plot area is divided into four quadrants marked I-IV as referred **[26]**. In fact, the four quadrants denote algorithm that achieves (i) higher S_n but lower S_p (ii) higher S_n and higher Sp (iii) lower S_n but higher Sp (iv) lower S_n and lower S_p. The diagonal line (0, 0) – (1, 1) denotes an algorithm that results in equal rates of true positives and false positives, i.e. a totally random method without any predictive power. Therefore, algorithms in quadrant II, far away from the diagonal line are better performers **[26]**.

All algorithms belonging to quadrant III tend to predict all the examples as negative thereby achieving high specificity but very low sensitivity. Aggrescan, Amylpred and FoldAmyloid appear in quadrant III indicating that although they have good Sp (scores of 86.15%, 91.06% and 90.84% respectively for AmylFibrilset and 85.4%, 90.5% and 91% respectively for Amylpreddataset), the S_n (scores of 22.45%, 12.67% and 16.33% respectively for AmylFibrilset and 26.2%, 20.7% and 16.6% respectively for Amylpreddataset), is poor. Out of these algorithms, FoldAmyloid achieves maximum FP rate with least TP rate for Amylpreddataset. Amylpred shows the highest FP rate but achieves lowest TP rate for AmylFibrilset. However, Amylpred scores the maximum MCC of .2 for Amylpreddataset.

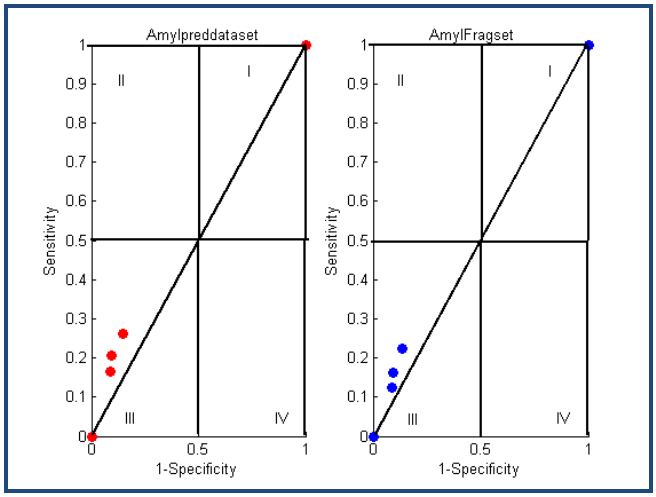

Figure 1: Scatter plot of True Positive rate (Sensitivity) versus false positive rate (1-Specificity) of three prediction tools for Amylpreddataset (red) and AmylFragset (blue).

It is also evident from **(Figure 1)** that there is not much significant difference in the statistical measures as far as the dataset is concerned. On further investigation, we observed that there were almost 20 positive hexmer examples which were predicted as fibril forming stretches by Aggrescan but not by other tools, resulting in the best TP rate. As a result, Aggrescan achieves significantly better BACC (score of .558).

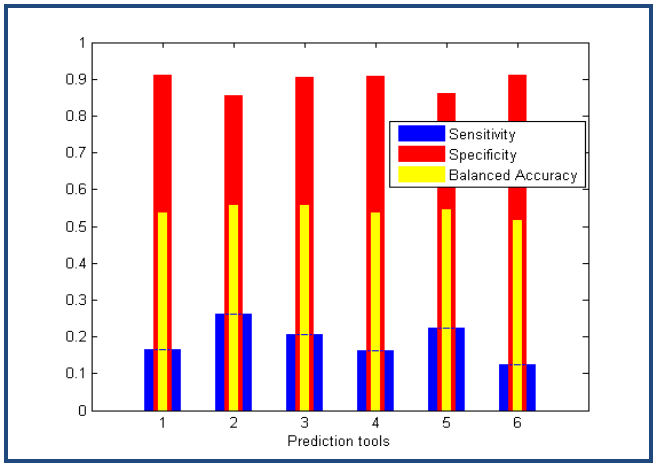

Figure 2: Comparative analysis of FoldAmyloid (1, 4 on X-axis), Aggrescan (2, 5 on X-axis) and Amylpred (3, 6 on X-axis) prediction tools on Amylpreddataset and AmylFibrilset respectively in terms of Sensitivity, Specificity and Balanced Accuracy.

The best quadrant of the plot in figure 1 is II with both S_n and s_p being > 0.5. As can be seen from the plot, none of the algorithms is found in this quadrant. However, the accuracy of these methods decreased significantly for Amylpreddataset/AmylFibrilset data and/or they suffered from highly biased prediction (very low S_n but very high S_p).

(Figure 2) depicts measures illustrating the rate of false positives and true positives, and the equilibrium maintained between the rates in terms of BACC. Table 2 shows a comparison of various methods under study in terms of the performance evaluation indicators namely TP, FN, FP, and TN along with S_n and S_p on Amylpreddataset and AmylFibrilset.

It remains difficult to assess accurately the performance of many of the tools listed above. In fact, experimentally mined amyloidogenic regions reported in different works do vary **[17]**. One possibility could be due to the fact that the protein sequences are examined under diverse states. Fibril formation depends on the experimental conditions and is expedited by denaturants: to aggregate, proteins should be unfolded, at least partly **[27]**. Hence reliable identification of amyloid fibril stretches is challenging and difficult.

Keeping in mind the various methodologies developed so far, one possible area of further research is to incorporate existing as well as new relevant features to develop efficient and effective algorithms for the same purpose. It was expected that Amylpred would result in better overall prediction accuracy than the other two tools for the reason that the performance of various unions and intersections of individual programs incorporated in this tool, lead to better predictions. Unfortunately, although it gives high S_p, the overall accuracy is poor due to very low S_n.

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 aggregation and implications for disease diagnosis and drug discovery. We have attempted to investigate the performance of few prediction tools in this study. To our knowledge, this is the first attempt to perform an evaluation on prediction tools in terms of prediction accuracy which remains as one of the key means to decipher the role of fibril forming regions in disease and therapeutics. However, the rapid development of computational methods for fibril forming prediction is promising for future research.

Recently published three methods of amyloid fibril forming segments identification tool have been compared to understand their relative performances. Statistical measures of these techniques on two datasets were measured. As evident, a significant reduction of S_n associated with a gain in Sp is noted. In other words, the tools got biased to predict most of the input instances as negative examples. However, of all the tools examined, Amylpred and FoldAmyloid showed the best S_p for AmylFibrilset and Amylpreddataset respectively, whereas Aggrescan displayed the maximum S_n irrespective of datasets. As far as the overall accuracy is concerned, certain improvements need to be incorporated in the prediction tools for a better performance.

Reference:

[1] Tian J *et al. BMC Bioinformatics*. 2009 **10**: S45 [PMID:19208147]
[2] Hamodrakas SJ, *FEBS J.* 2011 **278**: 2428 [PMID: 21569208].
[3] Pastor MT *et al. Curr. Opin. Struct. Biol*. 2005 **15**: 57 [PMID: 15718134].
[4] Monsellier E *et al. PLoS Comput. Biol*. 2008 **4**: e1000199 [PMID: 18927604].
[5] Caflisch A. *Curr. Opin. Chem. Biol*. 2006 **10**: 437 [PMID: 16880001].
[6] Gsponer J & Vendruscolo M, *Protein & Pept Lett*. 2006 **13**: 287 [PMID: 16515457].
[7] Nair SSK *et al.* IJCA Special Issue on "Computational Science - New Dimensions & Perspectives 2011.
[8] Thompson MJ *et al. Prac Natl Acad Sci U S A*. 2006 **103**: 4074 [PMID: 16537487].
[9] Zhang Z *et al. Bioinformatics*. 2007 **23**: 2218 [PMID: 17599928].
[10] Sole OC *et al. BMC Bioinformatics*. 2007 **8**: 65 [PMID: 17324296].
[11] Frousios KK *et al. BMC Struct Biol.* 2009 **9**: 44 [PMID: 19589171].
[12] López M & Serrano L, *Proc. Natl. Acad. Sci*. 2004 **101**: 87 [PMID: 14691246].
[13] Sanchez de Groot N *et al. BMC Struc. Biol*. 2005 **5**: 18 [PMID: 16197548].
[14] Castillo V & Ventura S, *PLoS Comput. Biol*. 2009 **5**: e1000476 [PMID: 19696882].
[15] Yoon S & Welsh WJ, *Protein Sci.* 2004 **13**: 2149 [PMID: 15273309].
[16] Galzitskaya OV *et al. PLoS Comput Biol*. 2006 **2**: e177 [PMID: 17196033].
[17] Nair SSK *et al. BMC Bioinformatics*. 2011 12(13): S21.
[18] http://bioinf.uab.es/aggrescan/
[19] Hamodrakas SJ *et al. Comput. Appl. Biosci*. 1998 4.
[20] Fernandez-Escamilla AM *et al. Nat Biotechnol*. 2004 **22**: 1302 [PMID: 15361882].
[21] http://biophysics.biol.uoa.gr/AMYLPRED/input.html
[22] Garbuzynskiy SO *et al. Bioinformatics*. 2010 **26**: 326 [PMID: 20019059].
[23] http://antares.protres.ru/fold-amyloid/
[24] Baldi P *et al. Bioinformatics*. 2000 **16**: 412 [PMID: 10871264].
[25] Levner I, *BMC Bioinformatics* 2005 **6**: 68 [PMID: 15788095].
[26] Bandyopadhyay S & Mitra R, *Bioinformatics*. 2009 **25**: 2625 [PMID: 19692556].
[27] Galzitskaya OV *et al. Mol Biol (MOSK)*. 2006 **40**: 5 [PMID: 17086993].

Edited by P Kangueane
Citation: Nair *et al.* Bioinformation 8(2): 070-074 (2012)

BIOINFORMATION

open access

Supplementary material:

Table 1: Methods and resources for amyloid fibril forming segments prediction

Tool	Resource	Method availability	Reference
Aggrescan	http://bioinf.uab.es/aggrescan/	Online search	Sole OC *et al.*, 2007 [18]
Amylpred	http://biophysics.biol.uoa.gr/AMYLPRED/input.html	Online search	Frousios KK *et al.*, 2009 [21]
FoldAmyloid	http://antares.protres.ru/fold-amyloid/	Online search	Garbuzynskiy SO *et al.*, 2010 [23]

Table 2: Prediction performance of each method in terms of the count of True Positives (TP), False Negatives (FN), False Positives (FP) and True Negatives (TN) along with Sensitivity (S_n) and Specificity (S_P) applied on AmylFibrilset [AF] (first three rows) and Amylpreddataset [AP] (last three rows) respectively

Tool	TP	FN	FP	TN	S_n	S_p
Aggrescan[AF]	264	912	127	790	.224	.861
Amylpred[AF]	149	1027	82	835	.126	.910
FoldAmyloid[AF]	192	984	84	833	.163	.908
Aggrescan[AP]	134	378	119	710	.262	.854
Amylpred[AP]	106	406	79	750	.207	.905
FoldAmyloid[AP]	85	427	74	754	.166	.910