Briefings in Bioinformatics, 22(6), 2021, 1–8

https://doi.org/10.1093/bib/bbab240
Review

Evaluation of *in silico* tools for the prediction of protein and peptide aggregation on diverse datasets

R. Prabakaran, Puneet Rawat, Sandeep Kumar and M. Michael Gromiha

Corresponding author. M. Michael Gromiha, Protein Bioinformatics Lab, Department of Biotechnology, Indian Institute of Technology Madras, Chennai, Tamil Nadu, India. Tel.: +91-4422574138; Fax: +91-4422574102; E-mail: gromiha@iitm.ac.in

Abstract

Several prediction algorithms and tools have been developed in the last two decades to predict protein and peptide aggregation. These *in silico* tools aid to predict the aggregation propensity and amyloidogenicity as well as the identification of aggregation-prone regions. Despite the immense interest in the field, it is of prime importance to systematically compare these algorithms for their performance. In this review, we have provided a rigorous performance analysis of nine prediction tools using a variety of assessments. The assessments were carried out on several non-redundant datasets ranging from hexapeptides to protein sequences as well as amyloidogenic antibody light chains to soluble protein sequences. Our analysis reveals the robustness of the current prediction tools and the scope for improvement in their predictive performances. Insights gained from this work provide critical guidance to the scientific community on advantages and limitations of different aggregation prediction methods and make informed decisions about their research needs.

Key words: protein aggregation; prediction; aggregation-prone regions; amyloid aggregation; tools

Introduction

Protein aggregation has been an active area research in recent years due to its association with human pathologies, a deteriorative role in the manufacturing, storage and shipping of protein therapeutics and enzymes for industrial applications;promise of viable mechanisms to build self-assembling nanostructures [1–6]. Amyloid fibrillization is associated with several human diseases such as Alzheimer's, Parkinson's, Huntington and various amyloidosis [7, 8]. Apart from the pathological importance, assessment of protein aggregation propensity is an essential part of the development of biotechnology products such as monoclonal antibodies, hormones, vaccines, enzymes and growth factors [9–12]. Higher aggregation propensity often correlates with low solubility and high viscosity, which affects product developability [13–15].

R. Prabakaran is currently doing integrated MS and PhD in Protein Bioinformatics at Indian Institute of Technology Madras. He has received his B.Tech Degree in Industrial Biotechnology from Anna University, India. His research is focused on understanding amyloidogenicity and aggregation propensity of human proteome.

Puneet Rawat, is working as a project scientist at Indian Institute of Technology, Madras, India. He has received his PhD in computational biology from Indian Institute of Technology, Madras, India. His research is focused on protein aggregation and computational immunology.

Sandeep Kumar, is a Distinguished Research Fellow (DRF) at the department of Biotherapeutics Discovery in Boehringer-Ingelheim Pharmaceutical Inc., Ridgefield, CT, USA. Sandeep Kumar holds a Ph.D. in Computational Biophysics and has over 20 years of experience researching protein structure — Function relationships using the computational means. Sandeep Kumar has so far contributed towards approximately 100 research articles, reviews, book chapters, and edited a book entitled "Developability of Biotherapeutics: Computational Approaches" published by CRC Press.

M. Michael Gromiha, received his PhD in Physics from Bharathidasan University, India and served as STA fellow, RIKEN researcher, research scientist and senior scientist at Computational Biology Research Center, AIST, Japan till 2010. Currently, he is working as a professor at Indian Institute of Technology Madras, India. His main research interests are structural analysis, prediction, folding and stability of globular and membrane proteins, protein interactions and development of bioinformatics databases and tools. He has published more than 250 research articles, 50 reviews, seven editorials, and two books entitled "Protein Bioinformatics: From Sequence to Function" by Elsevier/Academic Press and "Protein Interactions: Computational methods, analysis and prediction" by World Scientific.

Submitted: 8 April 2021; **Received (in revised form):** 18 May 2021

The growing interest and necessity to understand and control protein aggregation had triggered the development of several *in silico* tools during the last two decades [5, 16–18]. These various prediction tools can be classified based on input features to the model: protein sequence or structure and outputs obtained from the model such as β-aggregation propensity or amyloidogenicity, aggregation-prone regions (APRs) or β-sheet arrangement. Further, sequence-based features themselves can vary widely, ranging from sequence patterns; amino acid properties such as hydrophobicity, charge or secondary strand propensity; position-specific residue preferences and residue-pair preferences [12, 18–22].

However, the development and assessment of most of these prediction models are primarily based on a set of hexapeptide sequences [5, 20, 23–27] and extended datasets of experimentally identified APRs [28, 29]. Most of these methods have been well reviewed in the literature [5, 6, 18, 30]. In this work, we have performed a systematic comparative assessment of prediction performance using a series of diverse experimental datasets.

Methods and materials

Experimental datasets for performance comparison

Several diverse sequence datasets were collected from the literature to assess the prediction performance. These are described below and are available at https://web.iitm.ac.in/bioinfo2/fileShare/AggReview2/datasets/.

(i) **Hex1421**: Hex1421 consists of 512 amyloidogenic and 909 non-amyloidogenic hexapeptide sequences. These were collected from CPAD 2.0, WALTZ-DB and AmyLoad databases [26, 27, 31, 32]. Several prediction algorithms have been trained and validated on these hexapeptide sequences. Clustering the hexapeptides based on number of identical residues lead to 491, 26 and 1 cluster(s) at sequence identity cut-off of 33.33% (2 residues), 50% (3 residues) and 60.67% (4 residues), respectively. Further, normalized entropy-based conservation score calculated at each position showed that the residues were not conserved (score: 0.88–0.92; 0: conserved; 1: variable) [33, 34]. To avoid bias in performance comparison, sequences containing hexapeptides from Hex1421 were filtered out from other sequence datasets (listed below).

(ii) **Amyl37**: 37 amyloidogenic protein sequences with residue-level APR annotations were collected from AmyPro database [28] based on the following criteria: (a) sequence do not contain hexapeptides from Hex1421, (b) the protein contains less than 400 residues and (c) the APR fraction in the sequence is more than 5% and less than 100%. These restraints are added to filter out proteins with incomplete APR annotations.

(iii) **IAPP8**: Palato *et al*. [35] studied the amyloidogenicity of homologous sequences of islet amyloid polypeptide (IAPP) protein from 12 different species. Among the 12 sequences, 4 sequences (from Human, Dog, Seal and Rat) containing hexapeptides from Hex1421 were removed.

(iv) **ALBase678**: 678 sequences of immunoglobulin light chain variable regions along with their association with amyloidosis were obtained from ALBase database (http://albase.bumc.bu.edu/aldb). The dataset contains 177 amyloidogenic and 501 non-amyloidogenic variable light chain sequences, respectively. The sequence identity among these sequences is less than 90%, and sequences do not contain hexapeptides from Hex1421.

(v) **Fold-Agg77**: A dataset of 46 proteins whose folding kinetics have been well studied experimentally and 31 aggregating protein sequences [36–38]. The sequences share less than 90% sequence identity and do not contain exact matches from Hex1421.

(vi) **Sol2151**: A dataset of 1077 soluble and 1074 insoluble E. coli protein sequences, respectively, which were experimentally studied in the *in vitro* system by Niwa *et al*. [39]. The sequences share less than 90% sequence identity.

Selection of prediction methods/algorithms

Among the various prediction algorithms and tools, we selected nine prediction tools based on the following criteria: (a) the method takes amino acid sequence as input for prediction, (b) the tools accessible as a webserver or stand-alone application and (c) can process large datasets of >100 sequences in a single submission. The list of various APR and aggregation propensity prediction tools compared in this study (Table 1) is discussed below:

(i) AGGRESCAN is based on aggregation-propensity scale for 20 natural amino acids [40, 41] derived from *in vivo* mutagenesis experiments on Aβ fused to GFP. It predicts the intrinsic aggregation propensity of a given sequence and hotspots/APRs in protein sequences.

(ii) AgMata utilizes energy potentials obtained from a dataset of globular protein structures [42] and a pair of logistic regression models for predicting parallel and antiparallel beta-aggregation propensity of protein sequences.

(iii) AmyloGram employs n-gram analysis with reduction of amino acid alphabets to encode hexapeptides [43] and Random-forest classifier to predict amyloidogenic hexapeptides.

(iv) ANuPP is an ensemble of logistic regression models trained on clusters of hexapeptides using atom composition as input features [44].

(v) APPNN is a neural network-based method [45] using physicochemical and biochemical amino acid properties.

(vi) GAP uses statistical potentials derived from amino acid pair preferences in alternate and adjacent positions of amyloid fibril-forming and amorphous β-aggregating peptides, and Bayesian networks to compute the amyloid probability [21].

(vii) PASTA2 predicts β-aggregation based on energy functions for parallel and antiparallel beta-strands derived from a larger dataset of protein structures in comparison with AgMata [46].

(viii) TANGO predicts β-aggregation prone regions based on relative preferences for secondary structure conformations in the protein sequence [22], which is obtained from Boltzmann distribution of empirical/statistically derived energetic terms.

(ix) WALTZ is a combination of three position-specific sequence information: (a) scoring matrix derived from amyloidogenic hexapeptides, (b) helical, beta and solvation-related hydrophobicity propensities and (c) pseudoenergy matrix derived from *in silico* mutagenesis of polyalanine microcrystal structures [20].

Assessment of prediction performance

To assess the prediction performance of protein and peptide aggregation propensity, the total and average aggregation scores for each sequence were obtained from the nine selected aggregation tools. Receiver operating characteristic (ROC) curves were

Table 1. List of aggregation prediction tool/algorithms

Methods	Link	Reference
AGGRESCAN	http://bioinf.uab.es/aggrescan/	Conchillo-Solé *et al.* [40]
AgMata	https://bitbucket.org/bio2byte/agmata	Orlando *et al.* [42]
APPNN	http://cran.r-project.org/web/packages/appnn/index.html	Família *et al* [45]
AmyloGram	http://www.smorfland.uni.wroc.pl/shiny/AmyloGram/;http://github.com/michbur/AmyloGramAnalysis	Burdukiewicz *et al.* [43]
ANuPP	https://web.iitm.ac.in/bioinfo2/ANuPP/	Prabakaran *et al.* [44]
GAP	https://www.iitm.ac.in/bioinfo/GAP/	Thangakani *et al.* [21]
PASTA 2	http://old.protein.bio.unipd.it/pasta2/	Walsh *et al.* [46]
TANGO	http://tango.crg.es/	Fernandez-Escamilla *et al.* [22]
WALTZ	https://waltz.switchlab.org/	Maurer-Stroh *et al.* [20]

used to obtain the optimum threshold cut-off for every method in each dataset independently to avoid bias towards any single prediction algorithm. The optimum threshold for a dataset is that the threshold value corresponds to the upper-left corner in the ROC curve, calculated from min $((1-\mathrm{TPR})^2+\mathrm{FPR}^2)$. Area under the ROC curve (AUC) was used for performance comparison. In addition, the following measures were also used.

$$\text{True positive rate (Sensitivity, Recall)}, \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \quad (1)$$

$$\text{True negative rate (Specificity)}, \mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} = 1 - \text{false positive rate (FPR)} \quad (2)$$

$$\text{Accuracy, ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\text{Total sample size}} \quad (3)$$

$$\mathrm{Q-value}, \mathrm{Q} = \frac{\mathrm{TPR}+\mathrm{TNR}}{2} \quad (4)$$

$$\text{Matthews correlation coefficient, MCC} = \frac{(\mathrm{TP}*\mathrm{TN})-(\mathrm{FP}*\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \quad (5)$$

$$F1 \text{ score}, F1 = \frac{2\,\mathrm{TP}}{2\,\mathrm{TP}+\mathrm{FN}+\mathrm{FP}} \quad (6)$$

In the above equations, TP, TN, FP and FN stand for true positive, true negative, false positive and false negative, respectively. Bootstrap sampling ($n = 1000$), Student's *t*-test, and Mann–Whitney rank-sum test were performed using SciPy [47] to assess the statistical significance of the difference in performance. Segment OVerlap (SOV) scores: $\mathrm{SOV}_{\mathrm{APR}}$, $\mathrm{SOV}_{\mathrm{non\text{-}APR}}$, $\mathrm{SOV}_{\mathrm{average}}$ and $\mathrm{SOV}_{\mathrm{overall}}$ are used for the evaluation of identification of APR regions in protein sequences [48]. APR predictions are similar to secondary structure assignment. SOV scores the prediction performance based on the overlap between the predicted and actual and hence appropriate for APR prediction assessment.

Results and discussion

Prediction of peptide amyloidogenicity

The ability of peptide sequences to form amyloid-like fibrils has been extensively studied experimentally and computationally. Hex1421 consists of 512 amyloidogenic and 909 non-amyloidogenic hexapeptides from all experimentally studied sequences over the decades [26, 27, 32]. The dataset has been widely used in the literature as training and validation datasets.

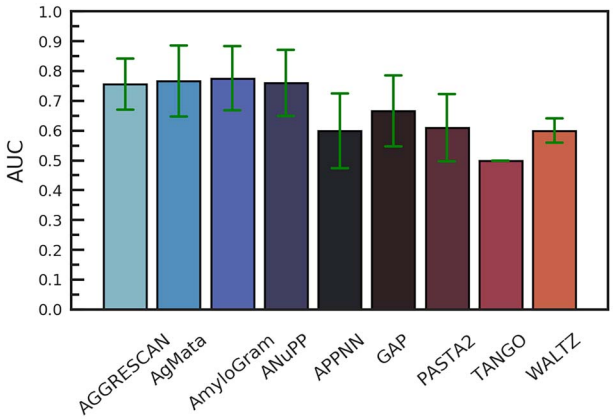

Figure 1. Performance (AUC) of prediction tools on the reduced hexapeptide dataset Hex31. The error bar indicates the standard deviation obtained from bootstrap sampling ($n = 1000$).

Seven out of the nine tools, namely AmyloGram, ANuPP, APPNN, GAP, PASTA2, TANGO and WALTZ have been either trained or validated using part of the Hex1421 dataset. Supplementary Table S1, available online at https://academic.oup.com/bib, shows the performance of the nine prediction tools on Hex1421 dataset. AmyloGram, ANuPP, AGGRESCAN and PASTA2 showed a performance with an AUC of greater than 0.8 [40, 43, 44, 46]. Further, to exclude the bias on the prediction performance, a dataset of 31 hexapeptides (**Hex31**) was constructed, which was not used for training in any of the methods. The methods AmyloGram, ANuPP, AGGRESCAN and AgMata scored an AUC above 0.75 (Figure 1) and the performance of AmyloGram, ANuPP and AGGRESCAN was consistently better than other methods on these hexapeptide datasets with a minimum accuracy of 75%.

Identification of aggregation prone regions

The performance of prediction tools was tested on a dataset of experimental APR annotation collected from the AmyPro database [28]. The dataset consists of annotation of 52 APRs in 37 amyloidogenic proteins. The assessments have been carried out using the measures, SOV score [48], the number of correctly predicted APR and non-APR segments. An APR/non-APR segment is considered correctly predicted if at least 50% of residues in a segment are correctly classified. AmyloGram was excluded in the analysis as it was trained to predict amyloidogenic and non-amyloidogenic peptides and not for predicting the aggregation prone regions in proteins.

Table 2. Performance of APR identification algorithms and tools

Methods	SOV_{APR}*	$SOV_{non\text{-}APR}$	$SOV_{overall}$	$SOV_{average}$	No. of correctly predicted APRs[a]	No. of correctly predicted non-APRs[a]
AGGRESCAN	58.2	48.2	46.9	53.2	28	43
AgMata	38.0	58.2	**53.8**	48.1	11	61
ANuPP	40.3	61.3	**57.0**	50.8	30	54
APPNN	58.9	36.4	36.6	47.6	32	28
GAP	39.3	26.0	27.2	32.7	16	26
PASTA2	54.8	44.6	44.6	49.7	31	41
TANGO	30.0	69.0	**60.9**	49.5	12	69
WALTZ	31.2	69.9	**62.0**	50.6	12	67

*SOV score was computed as in Zemla *et al*. [48]. [a]A segment (APR or non-APR) is counted as correctly predicted if more than 50% residues of the segment were identified by the method. In total, there are 52 APRs and 79 non-APRs in the Amy37 dataset. Methods, which showed a $SOV_{overall}$ of >0.5, are highlighted.
AUC values of top performing methods are highlighted in bold.

Overall, among 52 APRs, the maximum of 11–32 (21–61%) APRs are predicted correctly and SOV_{APR} and $SOV_{overall}$ ranged from 30 to 58.8% and 27.2 to 69.9%, respectively (Table 2). APPNN showed the highest SOV_{APR} of 58.9% and predicted 32 of the 52 APRs correctly. On the other hand, it showed a lower $SOV_{non\text{-}APR}$ of 36.4% and predicted only 28 of the 79 non-APR segments correctly. APPNN showed the highest sensitivity to APRs but also exhibited lower specificity. TANGO and WALTZ scored the highest $SOV_{overall}$ of 60.9 and 62%, respectively. The higher $SOV_{overall}$ is a result of higher $SOV_{non\text{-}APR}$ since the non-APRs are the majority class. Both TANGO and WALTZ showed higher specificity and lower sensitivity, respectively. As a result, they predicted 67 and 69 of 79 non-APRs correctly, respectively, but identified only 12 APRs. ANuPP, AGGRESCAN and PASTA2 showed a balanced sensitivity and specificity with $SOV_{overall}$ of 57, 47 and 44.6%, respectively. These three methods predicted 28–32 (54–60%) APRs and 41–54 (52–68%) non-APRs correctly. These results highlight the necessity of new robust methods to identify APRs accurately.

Amyloidogenicity of IAPP variants

IAPP is a 37-residue intrinsically disordered hormone protein known to form amyloid *in vitro* and *in vivo*. Interestingly, the amyloidogenicity of IAPP is known to vary among different species [49, 50]. In recent work, Palato *et al*. [35] studied the amyloidogenicity of 12 different naturally occurring full-length animal variants of IAPP by measuring the fibrillization using Thioflavin T binding and atomic force microscopy. Among 12 homologous protein sequences, 4 sequences which contained hexapeptides from Hex1421 were removed to avoid redundancy with the hexapeptide dataset. Figure 2 shows the predicted normalized aggregation scores (0–1) for the variants. Overall, most prediction methods showed excellent accuracy (>80%). AGGRESCAN showed the highest performance with a perfect ROC and AUC of 1 (Supplementary Table S2 available online at https://academic.oup.com/bib). AmyloGram, ANuPP, PASTA2, GAP and TANGO showed better performance with AUC in the range of 0.8–0.93 and an accuracy of 87.5%.

Amyloidogenicity of the immunoglobulin light chain variable regions

Light chain amyloidosis is the most common type of systemic amyloidosis and characterized by the deposition of immunoglobulin light chains in extracellular tissue [4, 51]. Amyloidogenicity of the light chains is attributed to the variable regions commonly observed in the tissue deposits [52]. A set of 678 light chain variable region sequences, including 356 kappa and 322 lambda subtypes, was obtained from the ALBase (http://albase.bumc.bu.edu/) (see Material and Methods section). These variable regions share high sequence similarity and can be clustered to four sequences at 40% sequence identity cut-off. Overall, we observed that all the methods showed poor performance in distinguishing the amyloidogenic and non-amyloidogenic light variable sequences (Table 3). The AUC varied from 0.374 to 0.674. Among the nine predictors: APPNN, ANuPP and AmyloGram showed better performance with AUC of 0.674, 0.606 and 0.579 and accuracy of 62.4, 59.2 and 59.8%, respectively.

Prediction of protein aggregating propensity

Fold-Agg77 consists of 46 proteins whose folding kinetics have been experimentally studied and 31 aggregating protein sequences [36–38]. Overall, the performance of methods as measured through AUC varied from 0.290 to 0.810. APPNN, TANGO and GAP showed the highest performance with an AUC of 0.810, 0.808 and 0.788 and accuracy of 79.3, 78.3 and 76%, respectively. AGGRESCAN, AmyloGram, ANuPP and PASTA2 showed similar performance with AUC between 0.720 and 0.743 and accuracy in the range of 70.5–74.2%.

Solubility and aggregation

Most often, aggregation is correlated with the solubility of proteins in solution. Proteins with higher aggregation propensity tend to separate from the bulk solution by coagulating or precipitating. Though the prediction of protein solubility is a significant problem, it is also a tricky one because of the subjectiveness of the definition of solubility itself for proteins. We tested the ability of aggregation prediction tools to distinguish the soluble and insoluble proteins using Ecoli2151 dataset [39]. These datasets were developed initially by Niwa *et al*. [39] and used previously by Agostini *et al*. [38] to develop the ccSOL. Overall, the AUC and accuracy varied in the range 0.568–0.777 and 55.9–71.7%, respectively (Table 5). PASTA2, APPNN, ANuPP and GAP showed better performance with an AUC of 0.777, 0.757, 0.747 and 0.740 and accuracy of 71.7, 70.1, 70.2 and 69.6%, respectively, than other methods. WALTZ and TANGO showed similar performance with AUC between 0.732 and 0.714 and accuracy in the range of 67.4–65.1%. This analysis highlights that the predicted aggregation scores could be used as a feature in the development of protein solubility prediction methods.

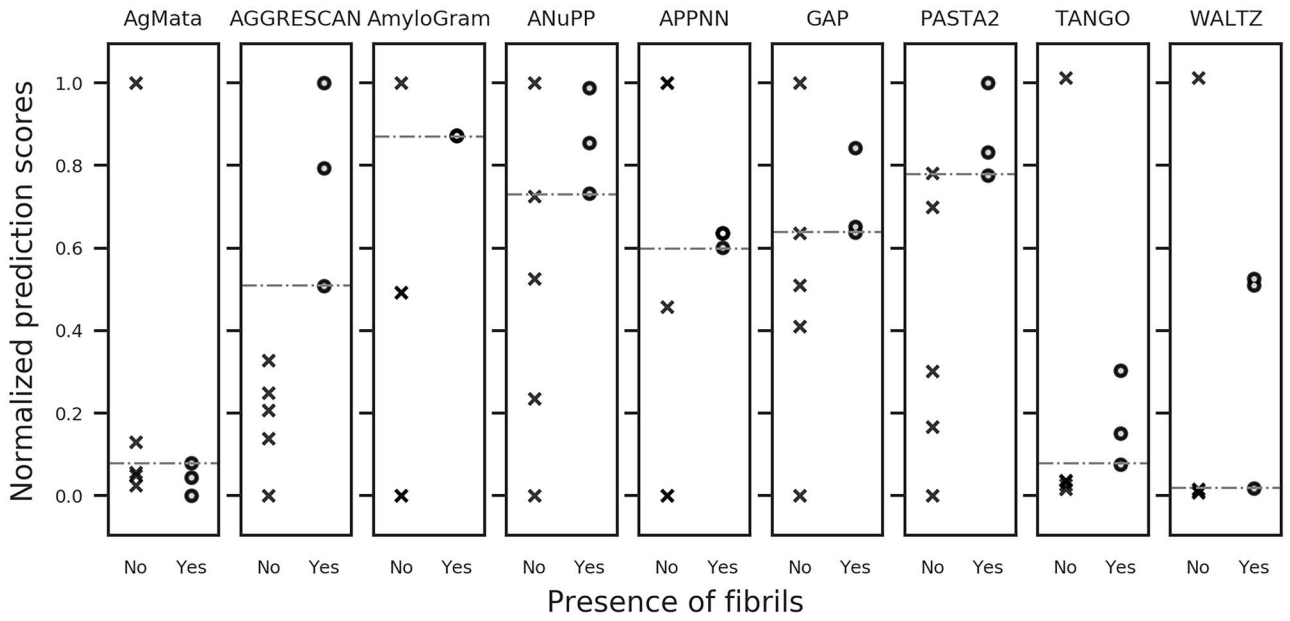

Figure 2. Prediction of diverse amyloidogenicity nature of the IAPP variants. The scatter plot shows the normalized predicted score (0–1) of the IAPP variants along with the experimental validation of the amyloidogenicity [35]. The circle and crosses indicate the presence and absence of amyloid fibrils, respectively. AGGRESCAN predicted score showed excellent correlation with the observed data with 100% accuracy (AUC = 1) in distinguishing the fibril forming and non-forming homologous proteins. ANuPP, TANGO, PASTA2, GAP and AmyloGram also scored accuracy greater than 85% (AUC≥0.8, Supplementary Table S2 available online at https://academic.oup.com/bib).

Table 3. Comparison of performance measures between different predictors on ALBase678 dataset

Methods	TPR (%)	TNR (%)	Accuracy (%)	Q-value (%)	AUC	F_1 score (%)	MCC
AGGRESCAN	41.4	45.5	44.4	43.4	0.374	27.5	-0.119
AgMata	70.1	34.4	43.8	52.3	0.444	39.4	0.043
AmyloGram	65.2	54.4	57.2	59.8	0.579	44.3	0.172
ANuPP	61.2	57.1	58.2	59.2	**0.606**	43.4	0.162
APPNN	70.6	59.5	62.4	65.1	**0.674**	49.5	0.265
GAP	54.3	41.4	44.7	47.8	0.426	33.7	-0.039
PASTA2	55.9	56.4	56.2	56.1	0.556	40.0	0.108
TANGO	43.2	52.9	50.3	48.0	0.447	31.0	-0.035
WALTZ	54.8	38.8	43.0	46.8	0.425	33.2	-0.059

The nine aggregation prediction tools were used to predict amyloidogenic light variable sequence from an experimental dataset of 678 annotated immunoglobulin light chain variable region sequences collected from ALBase (http://albase.bumc.bu.edu/).
AUC values of top performing methods are highlighted in bold.

Table 4. Comparison of performance measures between different predictors on Fold-Agg77 dataset

Methods	TPR (%)	TNR (%)	Accuracy (%)	Q-value (%)	AUC	F_1 score (%)	MCC
AGGRESCAN	70.7	71.9	71.4	71.3	0.720	66.2	0.421
AgMata	24.4	75.5	54.9	50	0.290	28.6	0.018
AmyloGram	74.5	67.8	70.5	71.2	0.723	66.7	0.416
ANuPP	69.4	72.5	71.3	71	0.743	65.7	0.416
APPNN	82.4	77.2	79.3	79.8	**0.810**	75.9	0.585
GAP	75.1	76.6	76.0	75.9	0.788	71.3	0.513
PASTA2	63.9	81.2	74.2	72.6	0.722	66.3	0.459
TANGO	80.4	77.0	78.3	78.7	**0.808**	74.6	0.564
WALTZ	72.3	50.8	59.5	61.6	0.571	58.0	0.240

The prediction performance was assessed on a dataset consists of 46 fast-folding protein whose folding kinetics have been well studied and 31 amyloid-forming protein sequences.
AUC values of top performing methods are highlighted in bold.

Table 5. Comparison of performance measures between different predictors on Sol2151 dataset

Methods	TPR (%)	TNR (%)	Accuracy (%)	Q-value (%)	AUC	F_1 score (%)	MCC
AGGRESCAN	65.5	62.2	63.9	63.9	0.688	64.4	0.278
AgMata	55.0	56.8	55.9	55.9	0.568	55.4	0.119
AmyloGram	64.2	56.8	60.5	60.5	0.635	61.8	0.210
ANuPP	75.2	65.2	70.2	70.2	**0.747**	71.6	0.407
APPNN	75.0	65.3	70.1	70.1	**0.757**	71.4	0.405
GAP	74.1	65.1	69.6	69.6	**0.740**	70.9	0.394
PASTA2	75.1	68.3	71.7	71.7	**0.777**	72.6	0.435
TANGO	62.2	67.9	65.1	65.1	0.714	64	0.303
WALTZ	64.8	70.1	67.5	67.4	0.732	66.5	0.350

Sol2151 dataset includes 1077 and 1074 soluble and insoluble E. coli protein sequences whose solubility was estimated experimentally using the *in vitro* expression system [39].
AUC values of top performing methods are highlighted in bold.

Table 6. Consolidated list of performance measures from the diverse benchmark datasets

	Hex1421	ALBase678	Fold-Agg77	Sol2151	Amy37	Avg. rank[a]	Avg. percentile[b]	Overall rank
	AUC				SOV overall			
AGGRESCAN	**0.838**	0.374	0.720	0.688	46.9	6.2	33.6	8
AgMata	0.407	0.444	0.290	0.568	53.8	7.4	18.9	9
AmyloGram	**0.886**	0.579	0.723	0.635	–	4.25	59.4	3
ANuPP	**0.854**	**0.606**	0.743	**0.747**	**57.0**	2.8	**76.8**	1
APPNN	0.753	**0.674**	**0.810**	**0.757**	36.6	3.2	70.3	2
GAP	0.697	0.426	0.788	**0.740**	27.2	5.6	40.0	6
PASTA2	**0.826**	0.556	0.722	0.777	44.6	4.2	58.2	4
TANGO	0.526	0.447	0.808	0.714	**60.9**	4.6	54.6	5
WALTZ	0.648	0.425	0.571	0.732	**62.0**	5.8	40.0	7

[a]Each predictor was ranked based on the listed performance scores and the average rank was calculated from the five datasets.
[b]To compare the prediction performance, scores were converted to percentile within each dataset, independently.
Scores of top performing methods in each dataset are highlighted in bold.

Overall comparison of prediction methods on diverse datasets

Methods, which showed good performance on diverse datasets (highlighted in Tables 2–5 and Supplementary Tables S1 and S2 available online at https://academic.oup.com/bib), had been compared in Table 6. The methods were ranked based on the predictions scores (AUC and SOV scores). Overall, ANuPP scored an average rank of 2.8 and a performance percentile of 76.8%. Other methods, APPNN, PASTA2 and TANGO ranked 3.2 (70.3%), 4.2 (58.2%) and 4.6 (54.6%), respectively.

The results presented in Table 6 showed that the performance of prediction methods varied across different datasets. For example, APPNN outperformed ANuPP on ALBase678, Fold-Agg77 and Sol2151 datasets. However, it underperformed in Hex1421, Hex31 and Amy37 datasets.

APPNN is a neural network model, trained on 296 hexapeptides and validated on a dataset of 483 peptides/protein sequences of varying lengths (>6 residues). The accuracy of the model in training and validation datasets are 78% [confidence interval (CI): 72.6–82.4%] and 84.9% [CI: 81.2–87.6%] [45]. We observed a similar trend in our analysis that APPNN performed better on protein sequences than hexapeptides. Moreover, the performance in the Hex1421 dataset is similar to the training dataset with an accuracy of 70.6% [CI: 68–73.1%].

AGGRESCAN, which is based on *in vivo* experiments on change in aggregation propensity among Amyloid-β mutants [41, 49], showed a moderate performance in all datasets except ALBase678. Though the method was not trained on hexapeptides, it showed an AUC of 0.838 in Hex1421 dataset. The observed prediction performance could be attributed to the amino acid level features of AGGRESCAN. ALBase678 is a challenging dataset of light chain variable sequences. These sequences are highly redundant with conserved framework regions and islands of variable CDR regions. ANuPP and APPNN showed the highest performance with an accuracy of 58.2 and 62.4% (AUC: 0.606 and 0.674), respectively. Both these methods were trained on hexapeptide datasets, and their robustness was tested on validation datasets as well as several cross-validation procedures. ANuPP is modeled with ensemble of logistic regression equations, whereas APPNN is based on artificial neural network. These results highlight the importance of large diverse datasets and robust model architecture to improve the performance of a method.

Ecoli2151 dataset is an extended inclusion in our analysis. Interestingly, four of the nine methods scored about 70% accuracy (AUC > 0.74), which indicates the inverse relationship between solubility and aggregation. Although AgMata and PASTA 2.0 were developed using beta-strand pairing scores, AgMata showed the least performance of 55.9% accuracy (AUC: 0.568). A similar trend was also observed in Fold-Agg77 dataset, where PASTA 2.0 (AUC = 0.722) showed better performance than AgMata (AUC = 0.29). The significant difference in these methods was mainly in the formulation and pair potentials used in these models [42, 53].

In summary, the performance of an aggregation prediction method depends on the model architecture, input features along with the training and validation datasets.

Conclusions

Several prediction algorithms and tools have been developed in the last two decades to predict beta-aggregation propensity and amyloidogenicity. Similar to experimental techniques, the development of *in silico* methods requires cycles of validation and refinement. In this current work, we have assessed the performance of nine different aggregation prediction tools on six diverse datasets. Some of these tools were developed to predict beta-aggregation in proteins, and others were trained to predict amyloid fibril formation. While some of these tools were trained on hexapeptide dataset, others were based on statistical analysis and scoring functions derived from sequence and structural dataset. Despite the difference, most methods showed good predictive power on the solubility dataset. Simultaneously, the ALBase dataset of highly similar immunoglobulin light chain variable regions proved to be a challenge for all the methods. Our analysis stresses the importance of benchmark datasets in the development of any tool. The current work provides a platform for further developing protein aggregation prediction tools to take advantage of the diverse datasets and suggests scope for improvement.

Key Points

- Systematically evaluated the performance of aggregation prediction tools on diverse curated datasets.
- Performance of a method depends on the type of datasets such as hexapeptides, protein sequences, amyloidogenic antibody light chains and soluble protein sequences.
- The method, ANuPP uniformly predicts well in all the datasets
- Provides insights to the scientific community to select a suitable method for prediction.

Supplementary data

Supplementary data are available online at *Briefings in Bioinformatics*.

Acknowledgements

We thank Bioinformatics Infrastructure facility, Department of Biotechnology and Indian Institute of Technology Madras for computational facilities and Ministry of human resource and development (MHRD) for HTRA scholarship to P.R. We would like to acknowledge the use of the Boston University ALBase, supported by HL68705, in this work.

Data Availability Statement

All the datasets used in the current work are available at https://web.iitm.ac.in/bioinfo2/fileShare/AggReview2/datasets/.

References

1. Pastor MT, Esteras-Chopo A, Serrano L. Hacking the code of amyloid formation: the amyloid stretch hypothesis. *Prion* 2007;**1**:9–14.
2. Kenney JM, Knight D, Wise MJ, *et al*. Amyloidogenic nature of spider silk. *Eur J Biochem* 2002;**269**:4159–63.
3. Linke RP, Glenner GG, Eanes ED, *et al*. Beta-pleated sheet fibrils. A comparison of native amyloid with synthetic protein fibrils. *J Histochem Cytochem* 1974;**22**:1141–58.
4. Dogan A. Amyloidosis: insights from proteomics. *Annu Rev Pathol Mech Dis* 2017;**12**:277–304.
5. Prabakaran R, Rawat P, Thangakani AM, *et al*. Protein aggregation: in silico algorithms and applications. *Biophys Rev* 2021;**13**:71–89.
6. Santos J, Pujols J, Pallarès I, *et al*. Computational prediction of protein aggregation: advances in proteomics, conformation-specific algorithms and biotechnological applications. *Comput Struct Biotechnol J* 2020;**18**:1403–13.
7. Chiti F, Dobson CM. Protein misfolding, amyloid formation, and human disease: a summary of progress over the last decade. *Annu Rev Biochem* 2017;**86**:27–68.
8. Dobson CM. Principles of protein folding, misfolding and aggregation. *Semin Cell Dev Biol* 2004;**15**:3–16.
9. Singla A, Bansal R, Joshi V, *et al*. Aggregation kinetics for IgG1-based monoclonal antibody therapeutics. *AAPS J* 2016;**18**:689–702.
10. Chennamsetty N, Voynov V, Kayser V, *et al*. Design of therapeutic proteins with enhanced stability. *Proc Natl Acad Sci U S A* 2009;**106**:11937–42.
11. Wang X, Singh SK, Kumar S. Potential aggregation-prone regions in complementarity-determining regions of antibodies and their contribution towards antigen recognition: a computational analysis. *Pharm Res* 2010;**27**:1512–29.
12. Thangakani A, Kumar S, Velmurugan D, *et al*. Distinct position-specific sequence features of hexa-peptides that form amyloid-fibrils: application to discriminate between amyloid fibril and amorphous β-aggregate forming peptide sequences. *BMC Bioinformatics* 2013;**14**:S6.
13. Kumar S, Thangakani AM, Nagarajan R, *et al*. Autoimmune responses to soluble aggregates of amyloidogenic proteins involved in neurodegenerative diseases: overlapping aggregation prone and autoimmunogenic regions. *Sci Rep* 2016;**6**:22258.
14. Nichols P, Li L, Kumar S, *et al*. Rational design of viscosity reducing mutants of a monoclonal antibody: hydrophobic versus electrostatic inter-molecular interactions. *MAbs* 2015;**7**:212–30.
15. Shan L, Mody N, Sormani P, *et al*. Developability assessment of engineered monoclonal antibody variants with a complex self-association behavior using complementary analytical and in silico tools. *Mol Pharm* 2018;**15**:5697–710.
16. Belli M, Ramazzotti M, Chiti F. Prediction of amyloid aggregation in vivo. *EMBO Rep* 2011;**12**:657–63.
17. Hamodrakas SJ, Liappa C, Iconomidou VA. Consensus prediction of amyloidogenic determinants in amyloid fibril-forming proteins. *Int J Biol Macromol* 2007;**41**:295–300.
18. Meric G, Robinson AS, Roberts CJ. Driving forces for nonnative protein aggregation and approaches to predict aggregation-prone regions. *Annu Rev Chem Biomol Eng* 2017;**8**:139–59.
19. Tartaglia GG, Cavalli A, Pellarin R, *et al*. Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. *Protein Sci* 2005;**14**:2723–34.
20. Maurer-Stroh S, Debulpaep M, Kuemmerer N, *et al*. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat Methods* 2010;**7**:237–42.
21. Thangakani AM, Kumar S, Nagarajan R, *et al*. GAP: towards almost 100 percent prediction for β-strand-mediated aggregating peptides with distinct morphologies. *Bioinformatics* 2014;**30**:1983–90.

22. Fernandez-Escamilla A-M, Rousseau F, Schymkowitz J, *et al*. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol* 2004;**22**:1302–6.
23. Louros N, Konstantoulea K, De Vleeschouwer M, *et al*. WALTZ-DB 2.0: an updated database containing structural information of experimentally determined amyloid-forming peptides. *Nucleic Acids Res* 2020;**48**:D389–93.
24. Tenidis K, Waldner M, Bernhagen J, *et al*. Identification of a penta- and hexapeptide of islet amyloid polypeptide (IAPP) with amyloidogenic and cytotoxic properties. *J Mol Biol* 2000;**295**:1055–71.
25. López De La Paz M, De Mori GMS, Serrano L, *et al*. Sequence dependence of amyloid fibril formation: insights from molecular dynamics simulations. *J Mol Biol* 2005;**349**:583–96.
26. Rawat P, Prabakaran R, Sakthivel R, *et al*. CPAD 2.0: a repository of curated experimental data on aggregating proteins and peptides. *Amyloid* 2020;**27**:128–33.
27. Wozniak PP, Kotulska M. AmyLoad: Website dedicated to amyloidogenic protein fragments. *Bioinformatics* 2015;**31**:3395–7.
28. Varadi M, De Baets G, Vranken WF, *et al*. AmyPro: a database of proteins with validated amyloidogenic regions. *Nucleic Acids Res* 2018;**46**:D387–92.
29. Tsolis AC, Papandreou NC, Iconomidou VA, *et al*. A consensus method for the prediction of 'aggregation-prone' peptides in globular proteins. *PLoS One* 2013;**8**:e54175.
30. Buck PM, Kumar S, Wang X, *et al*. Computational methods to predict therapeutic protein aggregation. *Methods Mol Biol* 2012;425–51.
31. Beerten J, Van Durme J, Gallardo R, *et al*. WALTZ-DB: a benchmark database of amyloidogenic hexapeptides. *Bioinformatics* 2014;**31**:1698–700.
32. Thangakani AM, Nagarajan R, Kumar S, *et al*. CPAD, curated protein aggregation database: a repository of manually curated experimental data on protein and peptide aggregation. *PLoS One* 2016;**11**:e0152949.
33. Shenkin PS, Erman B, Mastrandrea LD. Information-theoretical entropy as a measure of sequence variability. *Proteins* 1991;**11**:297–313.
34. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 1991;**9**:56–68.
35. Palato LM, Pilcher S, Oakes A, *et al*. Amyloidogenicity of naturally occurring full-length animal IAPP variants. *J Pept Sci* 2019;**25**:1–8.
36. Pawlicki S, Le Béchec A, Delamarche C. AMYPdb: a database dedicated to amyloid precursor proteins. *BMC Bioinformatics* 2008;**9**:273.
37. Tartaglia GG, Vendruscolo M. Proteome-level interplay between folding and aggregation propensities of proteins. *J Mol Biol* 2010;**402**:919–28.
38. Agostini F, Cirillo D, aria LCM, *et al*. ccSOL omics: a webserver for solubility prediction of endogenous and heterologous expression in *Escherichia coli*. *Bioinformatics* 2014;**30**:2975–7.
39. Niwa T, Ying B-W, Saito K, *et al*. Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. *Proc Natl Acad Sci* 2009;**106**:4201–6.
40. Conchillo-Solé O, de Groot NS, Avilés FX, *et al*. AGGRESCAN: a server for the prediction and evaluation of 'hot spots' of aggregation in polypeptides. *BMC Bioinformatics* 2007;**8**:65.
41. de Groot NS, Aviles FX, Vendrell J, *et al*. Mutagenesis of the central hydrophobic cluster in Abeta42 Alzheimer's peptide. Side-chain properties correlate with aggregation propensities. *FEBS J* 2006;**273**:658–68.
42. Orlando G, Silva A, Macedo-Ribeiro S, *et al*. Accurate prediction of protein beta-aggregation with generalized statistical potentials. *Bioinformatics* 2020;**36**:2076–81.
43. Burdukiewicz M, Sobczyk P, Rödiger S, *et al*. Amyloidogenic motifs revealed by n-gram analysis. *Sci Rep* 2017;**7**:12961.
44. Prabakaran R, Rawat P, Kumar S, *et al*. ANuPP: a versatile tool to predict aggregation nucleating regions in peptides and proteins. *J Mol Biol* 2020;166707.
45. Família C, Dennison SR, Quintas A, *et al*. Prediction of peptide and protein propensity for amyloid formation. *PLoS One* 2015;**10**:1–16.
46. Walsh I, Seno F, Tosatto SCE, *et al*. PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Res* 2014;**42**:W301–7.
47. Oliphant TE. Python for scientific computing. *Comput Sci Eng* 2007.
48. Zemla A, Venclovas Č, Fidelis K, *et al*. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins Struct Funct Genet* 1999;**34**:220–3.
49. Fox A, Snollaerts T, Errecart Casanova C, *et al*. Selection for nonamyloidogenic mutants of islet amyloid polypeptide (IAPP) identifies an extended region for amyloidogenicity. *Biochemistry* 2010;**49**:7783–9.
50. Alves NA, Frigori RB. In silico comparative study of human and porcine amylin. *J Phys Chem B* 2018;**122**:10714–21.
51. Hogan JJ, Lim MA, Dember LM. Light chain (AL) amyloidosis and the kidney. *Glomerulonephritis* 2018;1–10.
52. Blancas-Mejia LM, Misra P, Dick CJ, *et al*. Immunoglobulin light chain amyloid aggregation. *Chem Commun* 2018;**54**:10664–74.
53. Trovato A, Chiti F, Maritan A, *et al*. Insight into the structure of amyloid fibrils from the analysis of globular proteins. *PLoS Comput Biol* 2006;**2**:1608–18.