ANuPP: A Versatile Tool to Predict Aggregation Nucleating Regions in Peptides and Proteins

R. Prabakaran [1], Puneet Rawat [1], Sandeep Kumar [2]* and M. Michael Gromiha [1,3]*

*1 - **Protein Bioinformatics Lab,** Department of Biotechnology, Indian Institute of Technology Madras, Chennai, Tamil Nadu, India*
*2 - **Biotherapeutics Discovery,** Boehringer Ingelheim Pharmaceutical Inc., Ridgefield, CT, USA*
*3 - **School of Computing,** Institute of Innovative Research, Tokyo Institute of Technology, Yokohama, Kanagawa, Japan*

Correspondence to Sandeep Kumar, M. Michael Gromiha: *Sandeep_2.Kumar@Boehringer-Ingelheim.com (Sandeep Kumar), gromiha@iitm.ac.in (M. Michael Gromiha)*
https://doi.org/10.1016/j.jmb.2020.11.006
Edited by Michael Sternberg

Abstract

Short aggregation prone sequence motifs can trigger aggregation in peptide and protein sequences. Most algorithms developed so far to identify potential aggregation prone regions (APRs) use amino acid residue composition and/or sequence pattern features. In this work, we have investigated the importance of atomic-level characteristics rather than residue level to understand the initiation of aggregation in proteins and peptides. Using atomic-level features an ensemble-classifier, ANuPP has been developed to predict the aggregation-nucleating regions in peptides and proteins. In a dataset of 1279 hexapeptides, ANuPP achieved an area under the curve (AUC) of 0.831 with 77% accuracy on 10-fold cross-validation and an AUC of 0.883 with 83% accuracy in a blind test dataset of 142 hexapeptides. Further, it showed an average SOV of 48.7% on identifying APR regions in 37 proteins. The performance of ANuPP is better than other methods reported in the literature on both amyloidogenic hexapeptide prediction and APR identification. We have developed a web server for ANuPP and it is available at https://web.iitm.ac.in/bioinfo2/ANuPP/. Insights gained from this work demonstrate the importance of atomic and functional group characteristics towards diversity of atomic level origins as well as mechanisms of protein aggregation.

Introduction

Even though proteins have been continually evolving over a long time, aggregation remains a persistent threat to productive protein folding and function. This is evident from various protein aggregate depositions such as Lewy bodies and Amyloids implicated in many proteinopathies.[1] At the same time, various studies have also reported that aggregation propensities of proteins are related to their native state stability.[2–4] For example, Ma et al. [2] have shown that aggregation propensities of the subunit interfaces are greater than those of the molecular surfaces, and the differences between surface and interface aggregation propensities are directly correlated with optimal growth temperatures of organisms. By analyzing large datasets of protein structures, Buck et al. [4] and Prabakaran et al. [5] have shown that aggregation prone regions (APRs) in protein sequences contribute significantly more toward native state stability than other regions of similar lengths. Proteins have also evolved to control aggregation via negative design by interrupting the APRs through gate keeping residues.[6–7] Thangakani et al. [8] compared the incidence of APRs among thermophilic and mesophilic proteins, and showed that thermophilic proteins are better able to either interrupt APRs via gatekeeping residues or stow them in

their cores than their mesophilic homologues. The ability to control protein aggregation is also essential for the production of functional amyloid fibrils both *in vivo* and *in vitro* such as peptide hormone storage, scaffold support for melanin granules, biofilm formation and heritable information transfer.[9] Aggregation is a challenge for the design, development and storage of biotherapeutic molecules.[10]

Ability of the APRs to dictate the fates of proteins has attracted considerable research efforts. Understanding and accurate identification of the APRs has several potential applications in human diseases, development of biotherapeutics and design of nanomaterials.[11] Studies have shown that properties such as hydrophobicity, β-propensity, buriedness, and solvent isolatedness can identify APRs in protein sequences and structures.[4–5,12–13]

Most of the algorithms which have been developed to predict APRs in proteins are either sequence-based or structure-based mathematical models. The structure-based models such as SAP and Aggscore [14–15] quantified size and solvent exposure of the hydrophobic patches on the protein surface. Sequence-based methods, such as PAGE, TANGO, WALTZ and GAP, have highlighted the importance of sequence patterns, position-specific residue preferences and residue-pair preferences as well as charge, hydrophobicity and β-strand propensity to identify potential APRs in protein sequences.[13,16–18] Other methods such as Aggrescan3D and Solubis,[19–20] use a combination of both sequence and structure-based approaches.

While several algorithms have been devised to identify potential APRs in protein sequences and structures over the past couple of decades, no single algorithm currently succeeds consistently when used for aggregation prediction in diverse datasets. In fact, a recent comparative analysis of the existing methods has revealed that there is still a need for improvements in their predictive performances.[21] We postulate that use of amino acid residue-level features such as sequence patterns, residue preferences, and amino acid physicochemical properties limit the predictive power of these models because of the lack of large and diverse datasets of peptides and protein sequences currently available for training.

In this work, we have developed a new model called Aggregation Nucleation Prediction in Peptides and Proteins (ANuPP). ANuPP is an ensemble-classifier that identifies potential APRs in peptides and proteins by taking into account atomic-level features of hexapeptides. It showed an accuracy of 83% with an AUC of 0.883 in a test dataset of 142 hexapeptides and an average SOV of 48.7% for identifying APRs in 37 proteins. ANuPP is freely available for academic use at https://web.iitm.ac.in/bioinfo2/ANuPP/.

Methods

Datasets

ANuPP was trained on a dataset of hexapeptides that have been shown experimentally to be either amyloidogenic or non-amyloidogenic. Hexapeptides were chosen as the ideal length for the model because of availability of experimental data on a large number of them. Currently, 1421 experimentally validated amyloidogenic and non-amyloidogenic hexapeptides are available in literature.[22–26] In contrast, numbers of experimentally validated tri-, quadra-, penta- or hepta-peptides (lengths 3, 4, 5 or 7) are currently 1, 4, 16 and 50, respectively.[25] The hexapeptide dataset consists of 512 amyloidogenic and 909 non-amyloidogenic unique peptide sequences obtained from CPAD 2.0, WALTZ-DB, and AmyLoad databases.[22–26] The dataset was divided into two groups: 90% for training and cross-validation (**Hex1279)** and 10% as a blind test set for evaluating the performance of the method (**Hex142**).

Further, we validated ANuPP against APR annotations in AmyPro database.[27] We collected a dataset of 162 proteins from AmyPro. The dataset consists of annotations of identified Aggregation Prone Regions (APRs). However, we observed that the annotation of the APRs is incomplete and approximate. Several large proteins, such as AChE, BAP, CPEB, etc., are annotated with only one APR, and several APR segments are longer than 100 residues. The dataset was filtered to remove sequences without APR annotation, ambiguous annotations, and APR residue fraction is less than 10% or greater than 95%. Further, we performed clustering at 40% sequence identity using CD-HIT to reduce the redundancy and obtained a set of 54 proteins.[28] These protein sequences were analyzed for presence of hexapeptides from Hex1279, training dataset. Based on the number of hexapeptides present in the protein sequence, we divided the 54 proteins into Amy17 (>1 hexapeptide from Hex1279) and Amy37 (0 or 1 hexapeptide from Hex1279) datasets for calibration and assessment. The cut-off of 1 was chosen to avoid reducing the validation dataset size (Amy37). However, the choice of cut-off did not affect the conclusion of the analysis. The calibration set of 17 protein sequences are used to select an optimum value for parameters such as minimum nucleating window and threshold cut-offs (see "Identification of Aggregation Prone Regions"). Sequences in Amy17 and Amy37 share less than 40% sequence identity. All the datasets used in study are listed in the Appendixes (see Supplementary Information) and also available for download at https://web.iitm.ac.in/bioinfo2/ANuPP/datasets/.

Features design and selection

ANuPP uses atom compositions of peptides and protein segments as features for training the algorithm. Along with the size and branching of side chains, the presence or absence of specific functional groups underpins different physicochemical properties of the 20 amino acids (Supplementary Table S1). The use of atomic characteristics helps to expand the training feature sets. For example, atoms C^{β}, C^{γ}, $C^{\delta1}$, and $C^{\delta2}$ represent Leucine in the atom feature space. These atoms represent the presence of a side chain (C^{β}), length of the side chain, lack of non-aliphatic polar or charged functional groups, lack of branching at C^{β} and extension of the side chain by a methyl group (C^{γ}), presence of next-level branching at C^{γ} ($C^{\delta1}$ and $C^{\delta2}$), respectively. Since each amino acid contains multiple atomic level features, 1279 hexapeptide sequences in the training dataset contain several instances of occurrences for all the 36 atom types. Frequency of an atom type *i* is the number of atoms of type *i* normalized by the total number of atoms in a hexapeptide, as shown in equation (1) and illustrated in Supplementary Figure S1.

$$Frac_{atom\ i}(segment) = \frac{N_{atom\ i}(segment)}{\sum_{i=1}^{36} N_{atom\ i}(segment)} \quad (1)$$

PDB atom-naming convention[29] has been used to denote different atom types present in each amino acid residue. In this work, we focused only on 'heavy' atoms and excluded all hydrogen atoms. In all, 36 distinct atom types were defined to represent all twenty standard amino acids. Further, 15 out of 36 atom types, namely, $C^{\gamma1}$, $C^{\delta2}$, S^{δ}, $N^{\varepsilon2}$, C^{β}, $C^{\varepsilon2}$, OH, S^{γ}, $O^{\delta2}$, $C^{\varepsilon3}$, N^{ε}, $N^{\delta1}$, $C^{\varepsilon1}$, C^{ζ} and $C^{\eta2}$ were chosen to represent a given hexapeptide sequence via sequential forward feature selection. The selected features highlight diversity of physicochemical attributes of different atom types and functional groups that constitute each of the twenty naturally occurring amino acids. For example, $C^{\delta1}$ atom represents carbon atom at the delta position of branched amino acids such as Phe, Ile, Leu, Trp, and Tyr. $C^{\delta1}$ is present in both Ile and Leu; however, $C^{\delta1}$ and $C^{\delta2}$ are present in Leu and Ile, respectively. N^{ζ} and $N^{\eta1}$ together indicate uniquely the presence of positively charged residues: Arg and Lys, respectively.

Model architecture

ANuPP hypothesizes that aggregation mechanisms for amyloidogenic hexapeptide sequences vary depending upon their atomic compositions. For example, the isolation of hydrophobic side chains from the solvent would be the driving force for aggregation of the hexapeptide sequence rich in aliphatic and aromatic residues. On the other hand, hydrogen bonds and ion pairs may play active roles in the aggregation of a hexapeptide containing polar and charged residues. A few well-known examples include DFNKF from human calcitonin and GNNQQNY from yeast Sup35 protein.[30–31] The arrangement of Gln/Asn residues in the fibrils leads to the stacking of amide groups, which results in additional inter-strand hydrogen bonds.[32,33]

Supplementary Figure S2 shows the schematic representation of ANuPP architecture. ANuPP was developed as an ensemble-classifier to address the diversity in aggregation mechanisms. Initially, the 461 amyloidogenic hexapeptide sequences from the training dataset (Hex1279) were grouped into clusters using on K-means clustering based on physicochemical amino acid properties such as Hydrophobicity (H_p), total charge, number of charged residues, and total extended ASA (size) of the peptides. The 461 amyloidogenic sequences were clustered into various number of clusters (k) and an optimum k of nine clusters was selected such that each cluster contains a minimum of 20 sequences and have low Davies-Bouldin score and high silhouette score.[34–35]

An ensemble of nine independent logistic regression classifiers were built and trained on each of the nine clusters along with a randomly chosen subset of non-amyloidogenic hexapeptides (Supplementary Figure S2). To avoid class imbalance in the training dataset, we performed under-sampling, where the number of randomly chosen non-amyloidogenic hexapeptides was kept equal to the number of amyloidogenic hexapeptides. For a given segment, the predicted scores from the nine independent classifiers were combined using a Bayesian approach to derived the consensus score (Eqs. (2)–(4)).

$$P_{agg}{}^{cluster\ i}(segment) = \frac{1}{1 + e^{-\left(b_{i,0} + b_{i,1}x_1 + b_{i,2}x_2 + \cdots + b_{i,n}x_n\right)}}, \quad i = 1\ to\ 9, n = 1\ to\ 15 \quad (2)$$

$$P_{cluster\ i}(segment) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\left(\frac{x_i - \mu_i}{\sqrt{2}\sigma_i}\right)^2} \quad (3)$$

$$P_{agg}(segment) = \sum_{i=1}^{9} P_{cluster\ i}.P_{agg}{}^{cluster\ i} \quad (4)$$

In Eq. (2), P^{i}_{agg} represents the score for a given segment from the independent models, '*i*' varies from 1 to 9, $b_{i,n}$ represents the weight of the logistic regression model *i* for the n^{th} atom composition feature, '*n*' varies from 1 to 15. In Eq. (3), $P_{cluster}$ represents the probability of the segment belonging to the cluster; x_i is the distance of the sequence from the center of the cluster as measured using the physicochemical properties used for clustering, μ_i and σ_i are mean and standard deviation of the distance of the original cluster members from the center. P_{agg} is

the consensus score for the segment derived by combining the individual predicted scores in a Bayesian approach.

Identification of aggregation prone regions

To identify APR in protein sequences, a 4-step approach similar to Chou-Fasman secondary prediction algorithm was applied.[36] Initially, for a given sequence S of n residues is split into n-5 overlapping hexapeptide segments, and P_{agg} is predicted for each segment. Second, the residue aggregation score (P^{res}_{agg}) is calculated from the segment scores (P_{agg}) predicted by ANuPP. Aggregation score for a residue j in a sequence is the average of the predicted aggregation scores of all hexapeptide segments that contain the residue j. For example, aggregation score for residue number 5 is average of the predicted scores for hexapeptides spanning residues 1–6, 2–7, 3–8, 4–9 and 5–10. Following the residue score (P^{res}_{agg}) calculation, nucleating regions of minimum length ($w = 5$ residues) with average residue score ($<P^{res}_{agg}>$) higher than threshold 1 (th1 = 0.49) are identified across the sequence. These nucleating regions are further extended in N-terminal and C-terminal direction such that the residues have P^{res}_{agg} higher than threshold 2 (th2 = 0.31). The parameters involved in the prediction, i.e., minimum nucleating window length (w), threshold 1 (th1), and threshold 2 (th2) were optimized to score higher $SOV_{overall}$ on the calibration dataset of 17 AmyPro sequences (**Amy17)**.

Performance measures

The performance of the present method in hexapeptide datasets are assessed using the measures sensitivity, specificity, accuracy, Matthew's correlation coefficient and F_1 score.

$$\text{True Positive Rate (Sensitivity/Recall)}, TPR = \frac{TP}{TP + FN} \tag{5}$$

$$\text{True Negative Rate (Specificity)}, TNR = \frac{TN}{TN + FP} \tag{6}$$

$$\text{Accuracy}, ACC = \frac{TP + TN}{Total\ sample\ size} \tag{7}$$

$$\text{Q–value}, Q = \frac{TPR + TNR}{2} \tag{8}$$

$$\text{Matthews Correlation Coefficient}, MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{9}$$

$$F_1\text{score}, F1 = \frac{2TP}{2TP + FN + FP} \tag{10}$$

In the above equations, TP, TN, FP, and FN stands for true positive, true negative, false positive, and false negative, respectively. In addition, Receiver Operating Characteristic (ROC) curve and area under the curve (AUC) was computed to evaluate the performance. Further, the performance of ANuPP was compared with other predictors through Bootstrap sampling ($n = 1000$), Student's t-test, and Mann–Whitney ranksum test using SciPy.[37]

Segment OVerlap scores (SOV_{APR}, $SOV_{non\text{-}APR}$, and $SOV_{overall}$) are used to evaluate the performance of predicting APRs in proteins.[38] SOV scores the prediction accuracy based on the overlap between the predicted and actual segments instead of residues, which is more appropriate for segment prediction similar to secondary structure prediction. Further, exact residues in APRs are not well defined and hence prediction of APR regions is more reliable. In addition, $SOV_{average}$ was calculated as an average of SOV_{APR} and $SOV_{non\text{-}APR}$.

Results and Discussion

Development of a prediction model through ensemble approach

ANuPP was trained with Hex1279 dataset, containing 461 amyloidogenic and 818 non-amyloidogenic hexapeptides using an ensemble-based approach. It provides a robust framework to address diversity in amyloidogenic hexapeptides through independent model training. The performance of the method is presented in Table S2. We observed that ANuPP showed an accuracy, sensitivity and specificity of 79%, 68%, and 85%, respectively, with an Area under the ROC curve (AUC) of 0.852. Further, we have evaluated the performance using 10-fold cross-validation and the results are included in Table S2. ANuPP showed a consistent performance with an average sensitivity, specificity, accuracy and AUC of 63%, 86%, 77% and 0.831, respectively.

Assessment of ANuPP on the hexapeptide test dataset, Hex142

ANuPP was validated using a blind test dataset, Hex142, which contain 51 and 91 experimentally studied amyloid and non-amyloid hexapeptides, respectively. ANuPP showed a consistent performance with an AUC of 0.883 (Tables 1 and S3). Further, 83% of the hexapeptides are correctly classified as amyloids and non-amyloids with a sensitivity and specificity of 82% and 83%, respectively. To elucidate importance of the side chain functional group (atomic) features, we adapted the ANuPP model by re-training it with amino acid composition as input feature. The performance of the new model, $ANuPP_{AA}$ is listed in Table S2. $ANuPP_{AA}$ performed lower than ANuPP with an accuracy and AUC of 78% and 0.845, respectively in the test dataset. These

Table 1 Comparison of performances of different APR prediction methods on test dataset, Hex142.

	Sensitivity (%)	Specificity (%)	Accuracy (%)	Q (%)	F_1 score (%)	MCC	AUC
ANuPP	82.4	83.5	83.1	82.9	77.8	0.645	0.883
Aggrescan	68.6	85.7	79.6	77.2	70.7	0.551	0.855
FishAmyloid	45.1	82.4	69.0	63.8	51.1	0.296	0.798
GAP	94.1	27.5	51.4	60.8	58.2	0.26	0.721
Pasta2	37.3	96.7	75.4	67.0	52.1	0.45	0.855
TANGO	5.9	97.8	64.8	51.8	10.7	0.096	0.597
WALTZ	39.2	95.6	75.4	67.4	53.3	0.446	0.675

*AGGRESCAN[39]; Fish Amyloid[40]; GAP[13]; Pasta2[41]; TANGO[17]; WALTZ[18].

results highlight the contribution of atomic features to enhance the predictive power of the model.

The performance of our method is compared with other existing methods in the literature. Table 1 lists the performance of ANuPP against six existing predictors on the test dataset. ANuPP clearly showed the topmost performance with a Mathews Correlation Coefficient (MCC) of 0.645 and a balanced accuracy (*Q*) of 83%. Interestingly, our method showed a balance between sensitivity and specificity of 82% and 84%, respectively. GAP showed the highest sensitivity of 94% whereas its specificity dropped to 28%. An opposite trend was observed for PASTA2, TANGO and WALTZ with high specificity and low sensitivity. In addition, we used AUC for unbiased comparison to rank the predictors. We observed that ANuPP showed the highest AUC of 0.883 followed by PASTA2 and AGGRESCAN, which showed an AUC of 0.855 (Supplementary Figure S3). FishAmyloid scored an AUC of 0.798 with 69% accuracy.

Dependence of aggregation prediction on hydrophobicity

Hydrophobicity plays a key role in protein and peptide aggregation. To quantify the variations in prediction accuracy with hydrophobicity, the **Hex142**, test dataset was grouped into three classes based on number of hydrophobic residues (0–2, 3, 4–6 residues for dominance of polar/charged residues, equal number of hydrophobic and polar/charged residues and dominance of hydrophobic residues to represent the mechanisms of aggregation due to charged/polar and hydrophobic residues) containing 46, 46 and 56 hexapeptides, respectively. Following residues were considered as hydrophobic: Gly, Ala, Cys, Tyr, Val, Leu, Ile, Met, Phe and Trp.

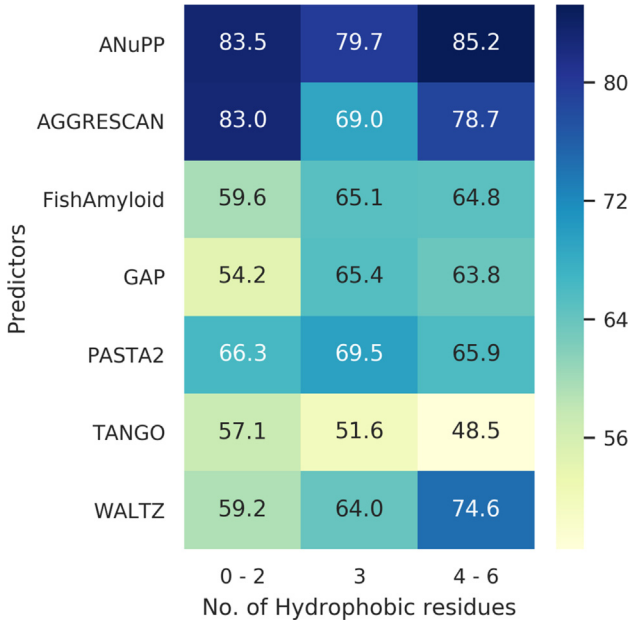

Figure 1. Variations in Balanced accuracy (Q) with hydrophobicity in Hex142, test dataset. The Hex142, test dataset was split into three classes based on number of hydrophobic residues (0–2, 3, 4–6 residues) containing 46, 46 and 56 hexapeptides, respectively. The balanced accuracy (Q) of different predictors was computed in each of these classes.

Table 2 Comparison of performance measures on the prediction of aggregation prone regions in amyloidogenic proteins based on Segment Overlap Score.

	SOV_{APR}	$SOV_{non-APR}$	$SOV_{Overall}$	$SOV_{Average}$	No. of correctly predicted APRs‡	No. of correctly predicted Non-APRs‡
ANuPP	45.2	52.3	50.2	48.7	28	50
Aggrescan	34.3	36.5	32.4	35.4	17	37
FishAmyloid	14.5	45.2	37.5	29.9	6	56
PASTA2 (85% specificity)	13.2	24.9	23.2	19.1	5	32
TANGO	19.1	57.8	48.1	38.5	7	69
WALTZ	44.4	28.9	28.7	36.6	25	28

*Segment Overlap score was computed as in Zemla et al. 38.
‡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 58 APRs and 81 non-APRs in the Amy37 dataset.

Figure 1 shows the performance of ANuPP along with the 6 other predictors over the three classes. ANuPP scored a balanced accuracy of greater 79% is all the three classes. In comparison, with the overall performance of 83% (Table 1), the difference was subtle. Though AGGRESCAN showed the second highest performance, there was a significant reduction in balanced accuracy (69%) in Class 2 hexapeptides with 3 hydrophobic residues in comparison with its overall accuracy of 77%.

Identification of aggregation prone regions

Further, we have validated the predictive power of our model to identify aggregation prone regions using the dataset of 37 amyloidogenic proteins collected from the AmyPro database, **Amy37** (see "Datasets"). We have evaluated the performance of the method using Segment Overlap scores and the results are presented in Table 2. SOVs for each class (APR and Non-APR) are calculated along with overall and average SOV to compare the performance of ANuPP with other existing methods. The results presented in Table 2 showed that ANuPP scored the highest overall and average SOV of 50.2% and 48.7%, respectively. TANGO showed the overall and average SOV of 48% and 39%, respectively. In addition, ANuPP showed a balance between SOV_{APR} (45.2%) and $SOV_{non-APR}$ (52.3%) scores whereas TANGO had SOV_{APR} and $SOV_{non-APR}$ scores of 19.1% and 57.8, respectively. On the other hand, Aggrescan showed a balanced performance with SOV_{APR} and $SOV_{non-APR}$ of 34% and 36%, respectively. Similar, imbalance was found in other methods except PASTA2.

ANuPP, prediction server and a repository

ANuPP web server is a user-friendly open platform built using Bootstrap, Django, and MySQL and runs on an apache server. For a given input sequence, ANuPP identifies potential APRs, lists aggregation score for every hexapeptide segment, and draws an interactive aggregation spectrum (Figure S4). The web server is available at https://web.iitm.ac.in/bioinfo2/ANuPP/.

ANuPP web server can accept up to 50,000 sequences in a single run to assist proteome-level analysis of aggregation propensity. ANuPP also acts as a repository as it hosts predictions for the human proteome and 64 million hexapeptide sequences. These predictions show that approximately 10,881,439 (17%) of the 64 million hexapeptide sequences are aggregation prone. Additionally, analysis of the human proteome using ANuPP identified 261,246 potential APRs in 42,357 human protein sequences (which includes 20,379 proteins and their isoforms). Interesting, only 2.4% of the 261,246 APRs were not flanked by gatekeeper residues (charged residues and proline). A repository of these predictions is available on our server.

Limitations of ANuPP

As stated earlier, a major limitation towards developing an accurate aggregation prone region prediction model is the limited availability of experimentally validated data. We have tried to overcome this problem by considering atomic-level chemical properties of different functional groups in hexapeptide sequences in the Hex1279 dataset. Availability of experimental data on a larger number of hexapeptides shall further improve ANuPP's predictive performance.

Conclusions

In this work, we have developed ANuPP, a web-based meta-classifier, to identify aggregation prone peptides and regions in proteins. The performance of ANuPP was evaluated using several different datasets, which demonstrate its superior predictive power and versatility. While several

APR prediction programs are currently available in literature, ANuPP is unique. It is the first sequence-based method that uses atom-based features and considers diversity of aggregation mechanisms. Results presented here provide credence to our hypothesis that aggregation in peptides and proteins originates at atomic rather than residue level.

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 PR. We thank WALTZ developers for sharing the executable.

Declaration of Competing Interest

Sandeep Kumar is an employee of Boehringer Ingelheim Pharmaceutical Inc. USA.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2020.11.006.

References

1. Chiti, F., Dobson, C.M., (2017). Protein misfolding, amyloid formation, and human disease: a summary of progress over the last decade. *Annu. Rev. Biochem.*, **86**, 27–68.
2. Ma, B.G. et al, (2010). Thermophilic adaptation of protein complexes inferred from proteomic homology modeling. *Structure*, **18** (7), 819–828.
3. Berezovsky, I.N., (2011). The diversity of physical forces and mechanisms in intermolecular interactions. *Phys. Biol.*, **8** (3) 2011 Jun; 035002.
4. Buck, P.M. et al, (2013). On the role of aggregation prone regions in protein evolution, stability, and enzymatic catalysis: insights from diverse analyses. *PLoS Comput. Biol.*, **9** e1003291.
5. Prabakaran, R. et al, (2017). Aggregation prone regions in human proteome: Insights from large-scale data analyses. *Proteins Struct. Funct. Bioinforma.*, **85**, 1099–1118.
6. Reumers, J. et al, (2009). Protein sequences encode safeguards against aggregation. *Hum. Mutat.*, **30** (3), 431–437. https://doi.org/10.1002/humu.20905.
7. Gsponer, J., Babu, M.M., (2012). Cellular strategies for regulating functional and nonfunctional protein aggregation. *Cell Rep.*, **2** (5), 1425–1437. https://doi.org/10.1016/j.celrep.2012.09.036.
8. Thangakani, A.M. et al, (2012). How do thermophilic proteins resist aggregation? *Proteins*, **80** (4), 1003–1015.
9. Invernizzi, G. et al, (2012). Protein aggregation: Mechanisms and functional consequences. *Int. J. Biochem. Cell Biol.*, **44**, 1541–1554.
10. Agrawal, N.J. et al, (2011). Aggregation in protein-based biotherapeutics: Computational studies and tools to identify aggregation-prone regions. *J. Pharm. Sci.*, **100**, 5081–5095.
11. Pastor, M.T. et al, (2007). Hacking the code of amyloid formation: the amyloid stretch hypothesis. *Prion*, **1**, 9–14.
12. Sawaya, M.R. et al, (2007). Atomic structures of amyloid cross-β spines reveal varied steric zippers. *Nature*, **447**, 453–457.
13. Thangakani, A.M. et al, (2014). GAP: towards almost 100 percent prediction for β-strand-mediated aggregating peptides with distinct morphologies. *Bioinformatics*, **30**, 1983–1990.
14. Sankar, K. et al, (2018). AggScore: prediction of aggregation-prone regions in proteins based on the distribution of surface patches. *Proteins Struct. Funct. Bioinforma.*, **86**, 1147–1156.
15. Chennamsetty, N. et al, (2009). Design of therapeutic proteins with enhanced stability. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 11937–11942.
16. Tartaglia, G.G. et al, (2005). Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. *Protein Sci.*, **14**, 2723–2734.
17. Fernandez-Escamilla, A.M. et al, (2004). Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.*, **22**, 1302–1306.
18. Maurer-Stroh, S. et al, (2010). Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat. Methods*, **7**, 237–242.
19. Van Durme, J. et al, (2016). Solubis: A webserver to reduce protein aggregation through mutation. *Protein Eng. Des. Sel.*, **29**, 285–289.
20. Zambrano, R. et al, (2015). AGGRESCAN3D (A3D): Server for prediction of aggregation properties of protein structures. *Nucleic Acids Res.*, **43**, W306–W313.
21. Prabakaran, R. et al, (2017). Influence of amino acid properties for characterizing amyloid peptides in human proteome. *Lect. Notes Comput. Sci.*, **10362**, 541–548.
22. Thangakani, A.M. et al, (2016). CPAD, curated protein aggregation database: a repository of manually curated experimental data on protein and peptide aggregation. *PLoS ONE*, **11** e0152949.
23. Beerten, J. et al, (2014). WALTZ-DB: A benchmark database of amyloidogenic hexapeptides. *Bioinformatics*, **31**, 1698–1700.
24. Wozniak, P.P., Kotulska, M., (2015). AmyLoad: Website dedicated to amyloidogenic protein fragments. *Bioinformatics*, **31**, 3395–3397.
25. Rawat, P. et al, (2020). CPAD 2.0: a repository of curated experimental data on aggregating proteins and peptides. *Amyloid*, **27**, 128–133.
26. Louros, N. et al, (2020). WALTZ-DB 2.0: an updated database containing structural information of experimentally determined amyloid-forming peptides. *Nucleic Acids Res.*, **48**, D389–D393.
27. Varadi, M. et al, (2018). AmyPro: a database of proteins with validated amyloidogenic regions. *Nucleic Acids Res.*, **46**, D387–D392.
28. Li, W., Godzik, A., (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
29. Burley, S.K. et al, (2019). Protein data bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.*, **47**, D520–D528.

30. Balbirnie, M. et al, (2001). An amyloid-forming peptide from the yeast prion Sup35 reveals a dehydrated β-sheet structure for amyloid. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 2375-.
31. Nelson, R. et al, (2005). Structure of the cross-β spine of amyloid-like fibrils. *Nature*, **435**, 773–778.
32. Bertolani, A. et al, (2017). Crystal structure of the DFNKF segment of human calcitonin unveils aromatic interactions between phenylalanines. *Chemistry*, **23**, 2051–2058.
33. Reddy, G. et al, (2010). Dry amyloid fibril assembly in a yeast prion peptide is mediated by long-lived structures containing water wires. *Proc. Natl. Acad. Sci.*, **107**, 21459–21464.
34. Rousseeuw, P.J., (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.
35. Davies, D.L., Bouldin, D.W., (1979). A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-1**, 224–227.
36. Chou, P.Y., Fasman, G.D., (1974). Prediction of protein conformation. *Biochemistry*, **13** (2), 222–245.
37. Oliphant, T.E., (2007). SciPy: Open source scientific tools for Python. *Comput. Sci. Eng.*, **9**, 10–20.
38. Zemla, A. et al, (1999). A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins Struct. Funct. Genet.*, **34**, 220–223.
39. Conchillo-Solé, O. et al, (2007). AGGRESCAN: a server for the prediction and evaluation of 'hot spots' of aggregation in polypeptides. *BMC Bioinf.*, **8**, 65.
40. Gasior, P., Kotulska, M., (2014). FISH amyloid – a new method for finding amyloidogenic segments in proteins based on site specific co-occurrence of amino acids. *BMC Bioinf.*, **15**, 54.
41. Walsh, I. et al, (2014). PASTA 2.0: An improved server for protein aggregation prediction. *Nucleic Acids Res.*, **42**, 301–307.