BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 14 2014, pages 1983–1990
doi:10.1093/bioinformatics/btu167

Sequence analysis

Advance Access publication March 28, 2014

GAP: towards almost 100 percent prediction for β-strand-mediated aggregating peptides with distinct morphologies

A. Mary Thangakani[1], Sandeep Kumar[2], R. Nagarajan[3], D. Velmurugan[1] and M. Michael Gromiha[3,*]

[1]Department of Crystallography and Biophysics, University of Madras, Chennai 600025, India, [2]Biotherapeutics Pharmaceutical Sciences, Pfizer Inc., Chesterfield, MO 63017, USA and [3]Department of Biotechnology, Indian Institute of Technology Madras, Chennai 600036, India

Associate Editor: John Hancock

ABSTRACT

Motivation: Distinguishing between amyloid fibril-forming and amorphous β-aggregating aggregation-prone regions (APRs) in proteins and peptides is crucial for designing novel biomaterials and improved aggregation inhibitors for biotechnological and therapeutic purposes.

Results: Adjacent and alternate position residue pairs in hexapeptides show distinct preferences for occurrence in amyloid fibrils and amorphous β-aggregates. These observations were converted into energy potentials that were, in turn, machine learned. The resulting tool, called Generalized Aggregation Proneness (GAP), could successfully distinguish between amyloid fibril-forming and amorphous β-aggregating hexapeptides with almost 100 percent accuracies in validation tests performed using non-redundant datasets.

Conclusion: Accuracies of the predictions made by GAP are significantly improved compared with other methods capable of predicting either general β-aggregation or amyloid fibril-forming APRs. This work demonstrates that amino acid side chains play important roles in determining the morphological fate of β-mediated aggregates formed by short peptides.

Availability and implementation: http://www.iitm.ac.in/bioinfo/GAP/.

Contact: gromiha@iitm.ac.in

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 20, 2014; revised on March 13, 2014; accepted on March 22, 2014

1 INTRODUCTION

Protein folding and aggregation are among the most important unsolved problems in biochemistry. In the past few decades, protein misfolding and aggregation have attracted much research interest not only because of their roles in conformational diseases, such as Alzheimer's and other amyloid neuropathies (Adams *et al.*, 2012; Checler and Turner, 2012; Liu *et al.*, 2012), but also because well-ordered aggregates can lead to the design of novel nanomaterials with desired characteristics (Cherny and Gazit, 2008; Knowles and Buehler, 2011; MacPhee and Dobson, 2000). Aggregation is also a major hurdle in the commercial manufacturing of biotechnology products such as enzymes and biopharmaceuticals (Agrawal *et al.*, 2011; Buck *et al.*, 2012; Kumar *et al.*, 2010). Aggregates formed by proteins and peptides exhibit a wide range of morphologies ranging from amorphous to well-defined amyloid fibrils with characteristic X-ray reflections (Eisenberg *et al.*, 2006; Sawaya *et al.*, 2007).

Formation of amyloid fibrils by short peptides is sequence dependent when all other parameters such as pH and concentration remain the same, indicating a role for side chains in the β-mediated aggregation process (Dobson, 1999; Lopez de la Paz and Serrano, 2004; Maurer-Stroh *et al.*, 2010). In a pioneering work from Serrano's laboratory, variants of a designed amyloid fibril-forming peptide, STVIIE, were tested for amyloid fibril formation at neutral and low pH (Lopez de la Paz and Serrano, 2004). This study provided valuable datasets of hexapeptides that form amyloid fibrils and those that form amorphous β-aggregates. The dataset of amyloid fibril-forming hexapeptides was expanded by the developers of WALTZ software, who derived position-specific matrices to further elucidate the rules for amyloid fibril formation in hexapeptides (Maurer-Stroh *et al.*, 2010). Moreover, several computational methods have been developed to identify the aggregation-prone regions (APRs) in proteins and peptides using descriptors such as β-strand propensity, hydrophobicity and charge (Belli *et al.*, 2011; de Groot *et al.*, 2012; Fernandez-Escamilla *et al.*, 2004; Maurer-Stroh *et al.*, 2010; Pawar *et al.*, 2005; Tjernberg *et al.*, 2002; Trovato *et al.*, 2007). Analyses of the available experimental data on hexapeptides that form amyloid fibrils (amyloid peptides) and those that form amorphous β-aggregates but not amyloid fibrils (amorphous β-aggregating peptides) enabled us to decipher rules that distinguish amyloid fibril hexapeptides from amorphous β-aggregating peptides (Thangakani *et al.*, 2013**).** It was observed that amyloid fibril hexapeptides have position-specific amino acid propensities that are distinct from those of the amorphous β-aggregating hexapeptides. By converting the position-dependent differences in amino acid propensities observed for the two types of hexapeptides into energy potentials, an algorithm to predict amyloid fibril peptides was devised. This algorithm compared well with the predictive levels of currently available algorithms (Kumar *et al.*, 2010). However, there was scope for further improvement.

In the present work, we extend our approach and probe propensities for various pairs of residues to occur simultaneously at alternate and adjacent position pairs in the β-strand hexapeptides that form amyloid fibrils or amorphous β-aggregates. We report the development of an algorithm [Generalized

*To whom correspondence should be addressed.

Aggregation Proneness (GAP)] that identifies both amyloid fibril and amorphous β-aggregate-forming hexapeptides with nearly 100 percent reliability (>96% accuracy). Furthermore, the performance of GAP was evaluated on five additional, but non-redundant, datasets that were not used in training this program. In all these datasets, GAP performed with significantly improved accuracy levels when compared with other commonly used methods for aggregation prediction. At a fundamental level, success of GAP demonstrates that propensities of amino acid residues to occur at adjacent and alternate position pairs may be crucial for determining the morphological fate of β-strand-mediated aggregates formed by short peptides. This observation complements the well-studied role of polypeptide backbone toward initiating aggregation in peptides and proteins (Chiti *et al*., 1999; Eakin *et al*., 2006). Besides accurately identifying potential APRs, GAP also predicts whether the APRs shall form amorphous β-aggregates or amyloid fibrils. This tool is expected to find several applications in both fundamental protein science and biotechnology. The three most important applications of GAP include the (i) ability to rationally design peptide-based nanomaterials with finely tuned morphologies and physical characteristics (Cherny and Gazit, 2008; Knowles and Buehler, 2011), (ii) improved identification of proteins in human and animal proteomes whose aggregation could lead to diseases and (iii) improved ability to identify different types of APRs in biopharmaceuticals and industrial enzymes (Agrawal *et al*., 2011; Buck *et al*., 2012; Kumar *et al*., 2010). Furthermore, it is anticipated that reliable identification of APRs in peptides and proteins shall facilitate design of novel aggregation inhibitors, for their use as therapeutic agents as well as biotechnological excipients, and protein engineering efforts to improve solubility via APR disruption (Agrawal *et al*., 2011; Buck *et al*., 2012; Kumar *et al*., 2010). Improved ability to identify APRs shall also enable fundamental research into relationships of aggregation with protein folding and function (Buck *et al*., 2013).

2 METHODS

2.1 Datasets of amyloid-forming peptides and non-amyloid peptides

In our earlier work (Thangakani *et al*., 2013), we have collected a set of 139 amyloid fibril-forming hexapeptides (Amyl139), called amyloid peptides for short, and 168 amorphous β-aggregating hexapeptides, called amorphous β-aggregating peptides (Amor168) (Lopez De La Paz and Serrano, 2004; Maurer-Stroh *et al*., 2010). These hexapeptides constitute our training set. The performance of our methods was tested on five different datasets:

(i) Hex40: 40 amyloid fibril-forming hexapeptides of non-identical sequences that are different from those in Amyl139.
(ii) Amyl310: 310 amyloid fibril-forming peptides with lengths 7–71.
(iii) AmylCryst15: 15 hexapeptides whose amyloid microcrystal structures are available from the Protein Data Bank (PDB) (Rose *et al*., 2013). The total number of hexapeptides with fibrillar crystal structures in the PDB is 24. However, 9 of these 24 hexapeptides are also present in the dataset of amyloid peptides used for training our method. These peptides were removed from AmylCryst for the sake of non-redundancy.
(iv) Twelve Sup35-derived peptides (Maurer-Stroh *et al*., 2010).
(v) Seventeen peptides derived from the N-terminal segment of huntingtin (Roland *et al*., 2013).

Hex40 and Amyl310 datasets were derived from the dataset used by Thangakani *et al*. (2012, 2013). Hex40 and AmylCryst15 do not have any sequences in common, either with each other or with the training dataset of 139 amyloid peptides. All the datasets used in this work are given in Supplementary Tables S1–S3.

2.2 Residue pair composition

Figure 1 shows a schematic diagram based on conformational details of a β-strand. A β-strand is a helical repeating pattern with two residues per turn (Creighton, 1993), implying that the residues that are next to each other in the amino acid sequence (K, K + 1) fall on alternate (opposite) faces of the helix, and the residues that are adjacent to each other on the same face of the helix are separated by a residue in the sequence (K, K + 2). For each alternate (K, K + 1; 1–2, 2–3, 3–4, 4–5 and 5–6) and adjacent (K, K + 2; 1–3, 2–4, 3–5 and 4–6) residue pairs in 139 amyloid (Amyl139) and 168 amorphous β-aggregating hexapeptides (Amor168), we have computed frequencies of residue pair types (i,j, where i and j are the 20 amino acids found in proteins) by using the following equation (Gromiha, 2010; Ou *et al*., 2013):

$$\mathrm{Comp(i,j)} = \Sigma n_{ij}/N \quad (1)$$

where $n_{i,j}$ is the number of residues of type i that are either adjacent or alternate to residues of type j, and N is the total number of hexapeptides (i.e. 139 for Amyl139 and 168 for Amor168). In total, nine different matrices were developed for amyloid fibrils accounting for each position-wise residue pair (alternate pairs, 1–2, 2–3, 3–4, 4–5 and 5–6; adjacent pairs, 1–3, 2–4, 3–5 and 4–6, see Fig. 1). Similarly, a separate set of nine different matrices was also developed for amorphous β-aggregating peptides. The total number of residue combinations is 400 (20 × 20) at each alternate or adjacent position pair.

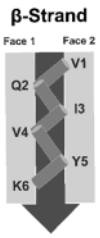

Fig. 1. Schematic diagram showing orientations of different residues in a β-strand. The residue pairs 1–2, 2–3, 3–4, 4–5 and 5–6 are at alternate faces (opposite) and the pairs 1–3, 2–4, 3–5 and 4–6 are at adjacent faces (same). Amyloid fibril and amorphous β-aggregating peptides show different alternate and adjacent residue pair preferences. These fundamental sequence features of hexapeptides determine the morphological fates of their β-mediated aggregates

2.3 Position pair-specific amino acid propensities

We have converted the composition of amino acid residue pairs at nine different pairs of positions of hexapeptides into propensities by normalizing them with an overall composition of residue pairs in globular proteins at adjacent and alternate positions (Thangakani *et al*., 2012, 2013). The propensity of amino acid residue pairs [Propen(i,j)] at different pair positions is given by

$$\mathrm{Propen(i,j)} = \mathrm{Comp(i,j)}/\mathrm{Compglob(i,j)} \quad (2)$$

where Compglob(i,j) is the composition of residues of type i that are at K, K + 1 or K, K + 2 sequence positions with residues of type j within a set of globular protein sequences (Gromiha, 2010; Gromiha *et al*., 2005).

2.4 Energy potentials

The residue pair propensities to occur at adjacent or alternate pair positions in Amyl139 and Amor168 peptides were treated as partition functions and converted into thermodynamic energy potentials [φ(i,j)] by using the following equation:

$$\Phi(i,j) = -RT \ln \text{Propen}(i,j) \quad (3)$$

where R is the universal gas constant, and T is the thermodynamic temperature. These energy potentials were used to obtain the potential difference for each residue pair type (i,j) to occur at alternate and adjacent positions in amyloid and amorphous β-aggregating hexapeptides using the following equation:

$$\Delta\phi(i,j) = \phi(i,j)_{\text{Amyloid}} - \phi(i,j)_{\text{Amorphous}} \quad (4)$$

2.5 Distinguishing between amyloid fibril-forming and amorphous β-aggregating hexapeptides using machine learning techniques

Several machine learning techniques, including Bayesian function, neural network, radial basis function network, logistic function, support vector machine, regression analysis, nearest neighbor, meta learning and decision tree and rules, implemented in WEKA (Witten and Frank, 2005) were tested for distinguishing between amyloid and amorphous β-aggregating hexapeptides. The details of all these methods are available in our earlier article (Gromiha and Suwa, 2006). We have used the energy potentials obtained from the difference between Amyl139 and Amor168 hexapeptides for the nine residue pairs as input features for these methods.

2.6 Assessment of predictive ability

We have performed 20-, 10- and 5-fold cross-validation tests for assessing the predictive ability of each machine learning technique used in the present work. In this procedure, the available dataset (139 amyloid and 168 amorphous β-aggregating hexapeptides) is divided into n groups, and $n-1$ of them are used for training a machine learning method. The rest is used for testing the method. The same procedure is repeated n times so that each datum is used at least once in the test.

We have used different measures, namely, sensitivity, specificity, accuracy and precision to assess the performance of machine learning methods toward distinguishing between Amyl139 and Amor168 peptides. Sensitivity shows the correct prediction of amyloid peptides, specificity is the correct prediction of amorphous β-aggregating peptides and accuracy indicates the overall assessment. These terms are defined as follows (Gromiha, 2010):

$$\text{Sensitivity} = TP/(TP + FN) \quad (5)$$

$$\text{Specificity} = TN/(TN + FP) \quad (6)$$

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN) \quad (7)$$

$$\text{Precision} = TP/(TP + FP) \quad (8)$$

where TP is amyloid peptides correctly predicted as amyloid peptides, FP is amorphous β-aggregating peptides incorrectly predicted as amyloid fibril-forming peptides, TN is amorphous β-aggregating peptides correctly predicted as amorphous β-aggregating peptides and FN is amyloid peptides incorrectly predicted as amorphous β-aggregating peptides. These also refer to the number of true positives, false positives, true negatives and false negatives, respectively. In addition, we have used the ROC (receiver-operating characteristic) curve to assess the performance of each machine learning technique used in this work. An ROC curve is a plot connecting the true-positive rate (sensitivity) and false-positive rate (1-specificity). It can be interpreted numerically using a parameter called AUC (area under the curve). There are several programs available in the literature that can perform ROC analyses and compute AUC (Sonego *et al.*, 2008).

2.7 Calculation of Z-score

Z-scores were calculated to test the significance of GAP scores for the 17 peptides given in Supplementary Table S2 (Roland *et al.*, 2013). For i^{th} peptide

$$\text{Z-score (i)} = (\text{GAPscore(i)} - \mu)/\sigma \quad (9)$$

where μ and σ are mean and standard deviations of GAP scores computed for the 17 peptides, respectively.

3 RESULTS

3.1 Alternate and adjacent residue pair compositions in amyloid fibril and amorphous β-aggregating peptides

We have computed alternate and adjacent residue pair propensities for all the hexapeptides in amyloid fibril (Amyl139) and amorphous β-aggregating (Amor168) datasets and converted them into statistical potentials (see Section 2). The preferred adjacent and alternate residue pairs in Amyl139 and Amor168 datasets are shown in Table 1. Some of the preferred residue pairs at different positions of Amyl139 and Amor168 hexapeptides are common (e.g. Ser–Thr at positions 1–2; Leu–Val at 2–3, Phe–Phe, Val–Ile/Leu at positions 3–4; Glu/Phe/Val–Ile and Ser–Phe at positions 3–5), indicating that these pairs may be important for initial 'homing in' of the self-associating β-strands that eventually nucleate aggregation. Besides these common residue pairs, several other preferred residue pairs are different between Amyl139 and Amor168. These different residue pairs may be crucial for deciding the morphological fate of the peptide aggregates and therefore can be used for distinguishing between amyloid fibril and amorphous β-aggregating peptides. The comparison of residue pair preferences in β-strands extracted

Table 1. Preferred residue pairs in amyloid and amorphous β-aggregating hexapeptides

Position pairs	Amyloid fibril	Amorphous β-aggregates
Alternate position pairs (on opposite faces of β-strands)		
1–2	GT, SF, **ST**	AE, KA, KT, **ST**
2–3	IV, **LV**, TF, TV, VF, YV	AE, EC, LF, **LV**, MF, TV
3–4	**FF**, FI, FN, LI, **VI, VL,** VN	EL, EN, **FF**, SL, **VI, VL**
4–5	**FI**, IF, II, LF, **LI, LY**, NF, **NI** NY, WI, YI	FF, **FI**, II, LF, **LI, LY**, **NI**
5–6	**IE, IF**, IM, **IQ, IS**, IT, IV	FF, FL, **IE, IF**, II, **IQ, IS**
Adjacent position pairs (on same faces of β-strands)		
1–3	GF, NV, **SV**, VF	KV, **SV**
2–4	IN, LI, QI, **TI**	AL, EL, **TI**
3–5	**EI, FI**, LY, **SF**, VF, **VI**, VY	**EI, FI, SF, VI**
4–6	**IE**, LV	**IE**, LL, LM

Note: Residue pairs that are preferred in both amyloid fibril-forming and amorphous β-aggregating hexapeptides are shown in bold.

from the globular proteins with Amyl139 (or Amor168) showed that the distribution of residue pairs to occur at alternate and adjacent position pairs is mostly uniform in the β-strands from globular proteins, whereas the hexapeptides in both Amyl139 and Amor168 datasets show distinct position–pair-specific preferences. Furthermore, the residue pair preferences are different between the β-strands in the globular proteins and the hexapeptide datasets and are used to parameterize GAP. For example, the hydrophobic residue pairs, Ala–Ala, Ala–Leu, Ala–Gly, Leu–Ala, Leu–Leu, Gly–Gly, Gly–Leu and Val–Ala, are highly preferred at the alternate faces of β-strands in globular proteins. These residue pairs are favored neither in the peptides contained in Amyl139 nor in Amor168 datasets. Conversely, the residue pairs that are preferred in Amyl139 or Amor168 are not favored in the β-strands from globular proteins. Further, we analyzed the preference of residue pairs in amyloids and disordered regions (Supplementary Table S4). We noticed that the residue pairs are dominated with charged and polar residues in disordered regions. In addition, Ala preferred to pair with other residues in all alternate and adjacent positions. The distinct preferences of residue pairs in disordered regions and amyloids could distinguish both peptides with an accuracy of 99%. Overall, these observations indicate that amino acid residue side chains play important roles in determining not only whether a solvent-exposed β-strand would aggregate but also the morphological fate of the aggregate.

3.2 Energy potentials derived from amino acid pair propensities

To distinguish between amyloid and amorphous β-aggregating hexapeptides, the energy potentials (ΔE) for all amino acid residue pairs (20×20 matrices) to occur at nine different pair positions (adjacent and alternate) in Amyl139 and Amor168 were computed using Equation (3) described in Section 2. The energy potential differences ($\Delta\Delta$E) for residue pairs to occur at a given position pair in amyloid and amorphous β-aggregating hexapeptides were also computed [Equation (4) in Section 2]. The preferred residue pairs (Supplementary Table S4) in Amyl139 are dominated by the combinations of both aliphatic/aromatic or aliphatic and aromatic residues. This is less prevalent in case of the Amor168 peptides (Supplementary Table S5). In both Amyl139 and Amor168, the numbers of preferred polar–polar, polar–charged and charged–charged residue pairs are small: one (Ser–Thr at position pairs 1–2) in Amyl139 and none in Amor168. These results are consistent with our earlier observations on distinct position-specific preferences between amyloid and amorphous β-aggregating hexapeptides (Thangakani *et al.*, 2013).

The difference in energy values for all the residue pairs plays a key role in distinguishing between amyloid and amorphous β-aggregating peptides, and these may be reflected in aggregation kinetic differences between the two types of hexapeptides. Moreover, aggregation may be triggered with the combination of two adjacent and/or alternate residue pairs, which is reflected in the energy difference $\Delta\Delta$E. Therefore, the use of $\Delta\Delta$E values for residue pairs at different positions could successfully distinguish between Amyl139 and Amor168 hexapeptides.

3.3 Machine learning can distinguish between amyloid and amorphous β-aggregating hexapeptides

We have used several machine learning techniques for distinguishing between Amyl139 and Amor168 hexapeptides, and the results are presented in Table 2. The parameters used in each machine learning technique are given in Supplementary Table S6. We have considered nine features based on the residue pairs and their locations in all the considered hexapeptides as input for training and cross-validation.

The performance was measured in terms of accuracy, sensitivity and specificity of the predictions (see Section 2). The Bayesian network-based method performed the best. It yielded a 10-fold cross-validation accuracy of 96% along with the highest sensitivity (135 peptides, 97.1%) obtained by using statistically derived position-specific residue pair energy potentials. We have also evaluated the performance using residue pair composition and specific pair position propensities. We observed similar results, and the energy-based potentials showed 2% higher accuracy than the other features. To further test the accuracy of GAP, additional datasets of amyloid fibril-forming datasets were used. In particular, a new dataset of 40 amyloid fibril-forming peptides (Hex40) was added to the cross-validation. Supplementary Table S7 shows the performance of GAP on these peptides with several measures of assessment. Our method could distinguish between amyloid fibril and amorphous β-aggregating hexapeptides with a 10-fold cross-validation accuracy of 93.7%. The sensitivity and specificity are 95.5 and 91.7%, respectively. GAP also correctly predicted 14 of the 15 hexapeptides in the AmylCryst15 dataset. Unfortunately, similar data on peptides that form amorphous β-aggregates are not available in the literature. Therefore, further testing of GAP is limited to amyloid fibril-forming peptides only.

3.4 Assessment using peptides of different lengths

GAP was tested on 310 amyloid fibril-forming peptides of different lengths (Amyl310). The following strategy was used to predict whether a query peptide is amyloid firbil forming: The peptide sequence is divided into six-residue-long windows that slide by one residue at a time. For each window, the scores for amyloid fibril formation and amorphous β-aggregation are

Table 2. Performance of different methods for discriminating between amyloid and amorphous β-aggregating hexapeptides

Method	Sensitivity	Specificity	Accuracy	AUC
Bayesian network	97.1	95.2	96.1	0.991
Neural network	91.4	97.0	94.5	0.984
Naive Bayes	97.1	91.1	93.8	0.992
RBF network	95.0	95.8	95.4	0.989
Support vector machines	95.7	97.0	96.4	0.964
k-nearest neighbor	89.2	95.8	92.8	0.933
Bagging	92.1	96.4	94.5	0.981
Decision table	89.9	91.5	91.5	0.964
J48 decision tree	89.2	95.2	92.5	0.944
Random Forest	92.1	95.2	93.8	0.980

computed using Bayesian network. These scores are summed for all the windows to obtain total scores for amyloid fibril formation and amorphous β aggregation for the whole peptide. The best of these two scores predicts the query peptide as either amyloid fibril forming or amorphous β-aggregating peptide. Results obtained with 310 peptides of different lengths are presented in Supplementary Figure SF1. The accuracy is 100% for most of the peptides. Specifically, all 14–20-residue-long peptides and those >22 residues are correctly predicted to be amyloid fibril-forming peptides. For most of the remaining peptides, the accuracy is >90%. In all these cases, GAP missed only one or two peptides for a given length. Overall, 302 of 310 (97.4%) peptides in Amyl310 dataset are correctly predicted to be amyloid fibril forming.

Four detailed examples of the performance of GAP on well-studied amyloid fibril-forming peptides of lengths, 20 (α-synuclein, 61–80), 31 (myoglobin, AB domain), 40 (Aβ peptide) and 71 [huntingtin (htt)$^{NT}Q_{42}P_{10}K_2$] residues (Giasson *et al*., 2001; Iconomidou *et al*., 2013; Pike *et al*., 1995) are shown in Figure 2. Relatively long peptide sequences were chosen here to demonstrate that GAP can work well with long peptide sequences also, even though it has been trained using data on short six-residue-long peptides. For each peptide, the probability of forming amyloid fibrils and amorphous β-aggregates are shown for all overlapping hexapeptides. Figure 2a shows the marked difference between the probabilities for amyloid and amorphous β-hexapeptide segments and average probability (0.998) for the whole α-synuclein peptide 61–80 favors amyloid fibril formation, in agreement with experimental observations. Both the peptides, Myoglobin and Aβ peptide, shown in Figure 2b and c have high probabilities for amyloid fibril-forming hexapeptide segments for most of their lengths. The average probabilities of forming amyloid fibrils by these peptides are 0.889 and 0.860, respectively. The 71-residue huntingtin peptide has the average probability of 0.948 for amyloid fibril formation, and Figure 2d shows that most of the overlapping hexapeptide segments prefer to form amyloid fibrils. These validation examples demonstrate that GAP recognizes APRs in peptide sequences with high fidelity.

3.5 Influence of specific residue pairs on predictive power of GAP

Use of propensities for different pairs of amino acid residues to occur together at alternate and adjacent faces of a β-strand in our method is a key difference between GAP and other methods currently available in the literature. Table 3 analyzes how residue pairs at specific positions contribute toward predictive power of GAP. Both types of residue pairs, alternate as well as adjacent, make important contributions toward determining the morphological fate of a peptide as amyloid fibril or amorphous β-aggregate. However, the alternate residue pairs are better at making the above distinction. The importance of each position pair was also evaluated (Table 3), and the core residue pairs (positions 2–3, 2–4, 3–4, 3–5 and 4–5) were found to be essential for distinguishing between amyloid fibril and amorphous β-aggregating peptides. However, using all position pairs together leads to significant improvement in the performance.

An attempt to combine position-specific single residue propensities from our previous work (Thangakani *et al*., 2013) and the

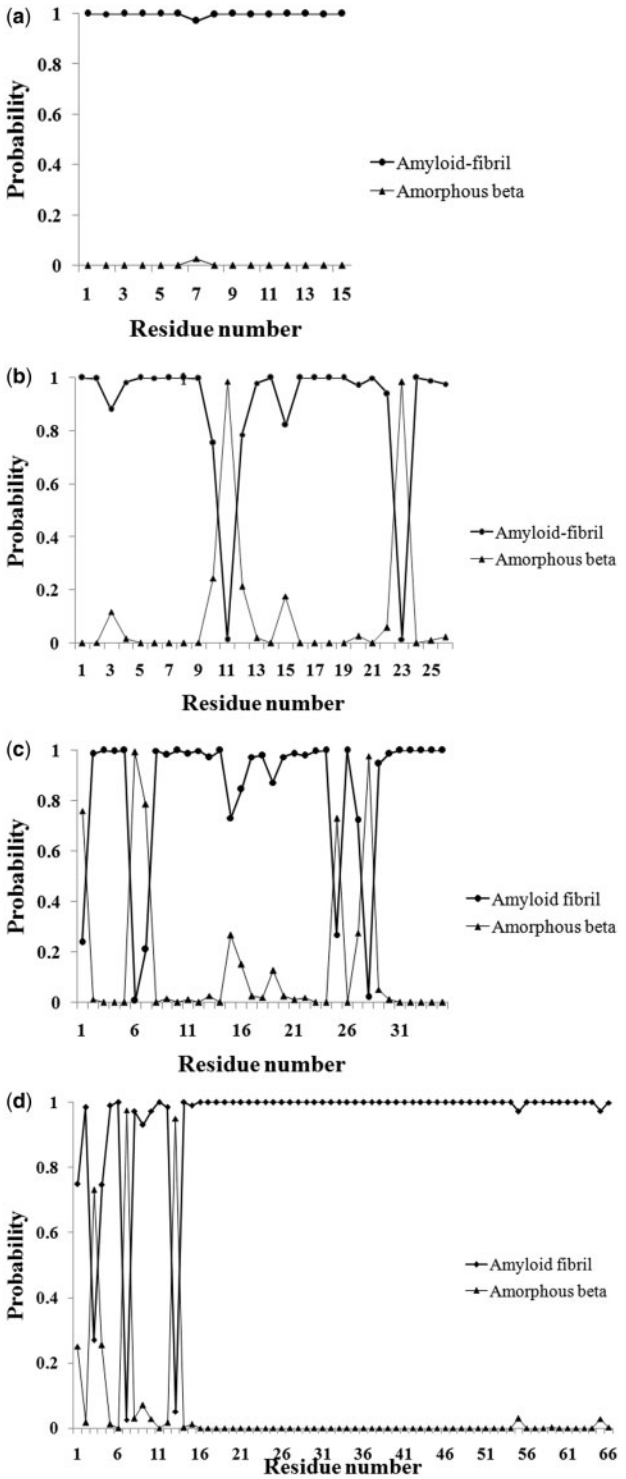

Fig. 2. Plots showing probabilities to form amyloid fibrils (black circle) and amorphous β-aggregates (black triangle) in four typical peptides of different lengths. The peptide sequences plotted here are: (**a**) α-synuclein, 61–80: EQVTNVGGAVVTGVTAVAQK; (**b**) myoglobin, AB domain: EGEWQLVLHVWAKVEADVAGHGQDILIRLFK; (**c**) Aβ peptide: DAEPRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVV; and (**d**) huntingtin (htt)$^{NT}Q_{42}P_{10}K_2$: TLEKLMKAFESLKSFQQQQ QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQP PPPPPPPPPKK

Table 3. Influence of residue pair positions on the ability of GAP to distinguish between amyloid fibril and amorphous β-aggregating hexapeptides

Residue pair position	Sensitivity	Specificity	Accuracy	AUC
1–2	89.9	83.7	77.2	0.820
2–3	95.0	63.1	79.5	0.806
3–4	86.6	72.6	79.8	0.780
4–5	88.8	59.5	74.6	0.718
5–6	87.7	64.9	76.7	0.767
All alternate	95.0	89.9	92.5	0.966
1–3	88.8	67.9	78.7	0.803
2–4	88.8	64.9	77.2	0.787
3–5	88.8	63.7	76.7	0.720
4–6	88.8	68.5	79.0	0.796
All adjacent	93.9	81.0	87.7	0.949
All pairs	97.1	95.2	96.1	0.991

Table 4. Performance comparison of GAP with other commonly used aggregation prediction methods

Method	Sensitivity or specificity (%)				
	Amyl139	Amor168	Hex40	AmylCryst15	Amyl310
PAGE0	98.6	0.6	100.0	100.0	99.4
PAGE5	87.8	8.3	97.5	100.0	94.8
AGGRESCAN	86.3	37.5	52.5	62.5	56.1
TANGO	22.3	92.3	2.5	6.7	71.6
WALTZ	81.3	77.6	7.5	6.7	47.1
Amylpred2	99.3	48.8	72.5	80.0	93.3
Previous work	79.9	83.9	32.5	46.7	42.6
Present work	97.1	95.2	90.0	93.3	97.4

Note: Specificity and sensitivity are given for Amor168 and all other datasets, respectively. PAGE0, aggregation score (lnP) of <0 is considered as amyloid; PAGE5, aggregation score (lnP) of <−5 is considered as amyloid.

residue pair propensities from this work was also made. This did not improve the performance, indicating that position-specific residue pair propensities may be optimum for the prediction of APRs.

3.6 Benchmarking the performance of GAP

Simultaneous parameterization for the formation of amorphous β-aggregates and amyloid fibrils is unique to GAP. As far as we know, no other method available in the literature can simultaneously distinguish between amyloid fibril-forming and amorphous β-aggregating peptides, except for the one developed earlier by our group (Thangakani *et al.*, 2013). This limits our ability to directly compare GAP with the other methods and benchmark its performance. Notwithstanding this limitation, comparing the performance of GAP with five commonly used methods, namely, PAGE, AGGREGSCAN, TANGO, WALTZ and Amylpred2 (Table 4), proved insightful. Here, correct identification of the peptides in Amyl139 and Amor168 datasets by different methods was used as benchmark for evaluating their performances. PAGE and AGGREGSCAN performed well at correctly identifying amyloid fibril-forming peptides in Amyl139 (122 and 120 hexapeptides, respectively), but not at correctly identifying peptides in the Amor168 dataset (14 and 63, respectively). On the other hand, TANGO correctly identified 155 (92.3%) of amorphous β-aggregating hexapeptides in Amor168, but only 31 (22.3%) of the hexapeptides in the Amyl139 dataset. The above observations can be rationalized on the basis of data used in parameterization of these programs (Conchillo-Solé *et al.*, 2007; Fernandez-Escamilla *et al.*, 2004; Tartaglia *et al.*, 2004). TANGO was developed to predict β-strand-mediated aggregation, in general, and is not specific to amyloid fibril morphology (Fernandez-Escamilla *et al.*, 2004; Maurer-Stroh *et al.*, 2010). The performance of WALTZ was found to be superior to other methods for both Amyl139 (81.3% correct predictions) and Amor168 (77.6% correct predictions). WALTZ uses position-specific matrices developed specifically for amyloid fibril morphology (see Supplementary Material in Maurer-Stroh *et al.*, 2010), and several peptides in Amyl139 were also used in the training set of WALTZ. Therefore, its high performance on Amyl139 dataset was expected. However, its high performance for the Amor168 dataset is surprising and shows that this method is able to correctly identify amorphous β-aggregating hexapeptides also. Our previous method using position-specific residue potentials of hexapeptides performed at par with WALTZ (Table 4). It correctly predicted 111 (of Amyl139) and 140 (of Amor168) hexapeptides as amyloid fibrils and amorphous β-aggregating, respectively. The second-generation consensus-based method, Amylpred2, correctly identified 138 (99.3%) of the 139 peptides in Amyl139, but only 48.8% (82) of the peptides in Amor168. In comparison with the above programs, GAP was able to correctly identify both amyloid fibril-forming amorphous β-aggregating hexapeptides with high sensitivity (97.1%) and specificity (95.2%).

To further benchmark the performance of GAP with other algorithms, the amyloid fibril-forming peptides from three different experimentally validated datasets, Hex40, AmylCryst15 and Amyl310 were also used (Table 4). The sequences contained in the above datasets did not form part of the training datasets for GAP. However, it is unknown to us whether these sequences were used in parameterization of the other programs. PAGE correctly identified all the peptides in Hex40, AmylCryst15 and Amyl310. TANGO and WALTZ correctly identified 71.6 and 47.1%, respectively, of the peptides in Amyl310. In the other two datasets, Hex40 and AmylCryst15, fewer than three (<10% of the peptides) haxapeptides were correctly identified by TANGO and WALTZ. Amylpred2 correctly predicted in the range of 59–80% in Hex40, AmylCryst15 and Amyl310 datasets (Table 4). Our earlier method predicted 13 and 7 peptides as amyloid fibrils in Hex40 and AmylCryst15, respectively. In comparison with the above methods, GAP consistently performed at the accuracy levels of ≥90% (Table 4).

GAP was also tested on 48 experimentally determined amyloid fibril-forming peptide segments of different lengths from 33 well-known amyloidogenic proteins (Tsolis *et al.*, 2013). It correctly predicted 47 (98%) of them (Supplementary Table S8). Further, developers of WALTZ (Maurer-Stroh *et al.*, 2010)

had benchmarked the performance of their method by using 12 Sup35-derived 10-residue-long peptides that were shown form amyloid fibrils experimentally. The performance of GAP along with other prediction methods TANGO, WALTZ and Amylpred2 is shown in Supplementary Table S1. GAP correctly predicted all the 12 peptides (100% sensitivity). These peptides were not part of training set for GAP. Previously, the maximum sensitivity of 67% was attained by WALTZ.

Recently, Roland *et al.* (2013) have tested abilities of four aggregation prediction methods, namely, Zyggregator, TANGO, WALTZ and Zipper using a dataset of 15 scrambled peptide sequences derived from 18-residue-long N-terminal segment of huntingtin ($htt^{NT}Q$) that forms α-helix-rich aggregates under physiological conditions. They report that 5 of the 15 peptides form β-rich amyloid fibrils under the nearly physiological conditions described in the report. In addition, two more peptides were found by the authors to form amyloid fibrils at higher concentrations. The Lys6Ala mutant of the parent peptide ($htt^{NT}Q_{K6A}$) was also reported to form amyloid fibrils. Overall, Roland *et al.* (2013) studied 17 peptides, and 8 of these form amyloid fibrils, whereas the remaining 9 do not form any detectable aggregates under the reported conditions. These peptides were again not part of training set for GAP. GAP is able to distinguish between amyloid fibril-forming peptides and non-aggregating peptides, but not perfectly. Other algorithms had also performed poorly on this set of peptides [see Table 1 in Roland *et al.* (2013)]. Supplementary Table S2 and Supplementary Figure SF2 show the results obtained from the predictions made by GAP on this set of 17 peptides. For each peptide, the GAP score was computed as the average of amyloid fibril-forming probabilities for all the overlapping hexapeptide segments of the peptide obtained by sliding a window of six residues, one residue at a time, along the peptide sequence. The average GAP score for 17 peptides in the study by Roland *et al.* (2013) is 0.735 ± 0.087 (range, 0.563–0.932). For the eight peptides that form amyloid fibrils, the average GAP score is higher at 0.772 ± 0.084 (range, 0.684–0.932). On the other hand, the average GAP score for the 9 peptides that do not aggregate is lower (0.702 ± 0.081; range, 0.563–0.797) than the average for all 17 peptides. Supplementary Figure SF2 plots the Z-scores obtained using Equation (9) in Section 2 for all the 17 peptides. A line at zero Z-score is drawn in this figure. It can be seen that five of eight amyloid fibril-forming peptides fall above the line, that is, their GAP scores are more than the average GAP score for this set of peptides. Similarly, five of nine non-aggregating peptides fall below this line. It should be noted here that Roland *et al.* (2013) performed these experiments at near-physiological conditions. GAP was trained on hexapeptides for which the experiments were performed under non-physiological conditions (Lopez de la Paz and Serrano, 2004; Maurer-Stroh *et al.*, 2010; Thangakani *et al.*, 2013). Comparatively low peptide concentrations in the experiments by Roland *et al.* (2013) could be another factor.

3.7 Limitations of GAP and future directions

The benchmarking studies described above show that despite substantial improvements in predictive power of GAP, there are limitations to GAP. The experimental data used in the parameterization of this program (Lopez de la Paz and Serrano, 2004; Maurer-Stroh *et al.*, 2010) were restricted to measurements at the single peptide concentrations at acidic and neutral pH. Hence, although our method is able to discriminate between amyloid fibril-forming and amorphous β-aggregating peptides to high degrees of sensitivity and specificity, it is not yet able to predict the effect of concentration and pH on aggregation behavior of the hexapeptides. The method was developed with a small dataset of 139 amyloids and 169 amorphous peptides and shall need further refinements when additional data become available. Another potential improvements to GAP shall be the ability to predict for a given hexapeptide region, its propensity to form a β-strand, amorphous β-aggregates and amyloid fibrils in a step-wise manner. Further, the work on the effect of mutations on aggregation propensity of peptide and protein sequences is also in progress.

4 CONCLUSION

The amino acid sequences of amyloid fibril-forming and amorphous β-aggregating hexapeptides show distinct residue pair preferences to occur at adjacent and alternate positions. This observation enabled us to derive statistical potentials for amino acid residue pairs to simultaneously occur at the alternate and adjacent position pairs in hexapeptides that have been experimentally shown in the literature to from either amyloid fibrils or amorphous β-aggregates. These residue pair preferences have been used to develop a new tool called GAP. GAP is capable of accurately identifying APRs in protein and peptide sequences and also distinguishing whether the predicted APRs shall predominantly form amyloid fibrils or amorphous β-aggregates. This work also improved our fundamental understanding of β-strand-mediated aggregation process by highlighting the role of side chains in determining the morphological fate of the aggregates.

ACKNOWLEDGEMENTS

A.M.T. and D.V. thank the Bioinformatics Facility of the University of Madras for computational facilities. S.K. acknowledges his discussions with Drs Patrick Buck and Satish Singh on topics related to aggregation in proteins and peptides. Patrick Buck is thanked for providing us the PDB codes for the amyloid peptide crystals. M.M.G. wishes to thank Bioinformatics facility and IIT Madras for infrastructure facilities.

Funding: Department of Biotechnology, Government of India (BT/PR7150/BID/7/424/2012) (partially).

Conflict of Interest: none declared.

REFERENCES

Adams,D. *et al.* (2012) Amyloid neuropathies. *Curr. Opin. Neurol.*, **25**, 564–572.

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.

Belli,M. *et al.* (2011) Prediction of amyloid aggregation *in vivo*. *EMBO Rep.*, **12**, 657–663.

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.

Buck,P.M. *et al.* (2012) Computational methods to predict of aggregation in therapeutic proteins. In: Voynov,V. and Caravella,J. (eds) *Therapeutic Proteins: Methods and Protocols, Methods in Molecular Biology*. Vol. 899, Chapter 26. Springer, USA, pp. 425–451.

Checler,F. and Turner,A.J. (2012) Journal of Neurochemistry special issue on Alzheimer's disease: 'amyloid cascade hypothesis–20 years on. *J. Neurochem.*, **120** (**Suppl. 1**), iii–iv.

Cherny,I. and Gazit,E. (2008) Amyloids: not only pathological agents but also ordered nanomaterials. *Angew. Chem. Int. Ed. Engl.*, **47**, 4062–4069.

Chiti,F. *et al.* (1999) Designing conditions for *in vitro* formation of amyloid protofilaments and fibrils. *Proc. Natl Acad. Sci. USA*, **96**, 3590–3594.

Conchillo-Solé,O. *et al.* (2007) AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics*, **8**, 65.

Creighton,T.E. (1993) *Proteins: Structure and Molecular Properties*. 2nd edn. W.H. Freeman and Company, New York, pp. 186–187.

de Groot,N.S. *et al.* (2012) AGGRESCAN: method, application, and perspectives for drug design. *Methods Mol. Biol.*, **819**, 199–220.

Dobson,C.M. (1999) Protein misfolding, evolution and disease. *Trends Biochem. Sci.*, **24**, 329–332.

Eakin,C.M. *et al.* (2006) A native to amyloidogenic transition regulated by a backbone trigger. *Nat. Struct. Mol. Biol.*, **13**, 202–208.

Eisenberg,D. *et al.* (2006) The structural biology of protein aggregation diseases: fundamental questions and some answers. *Acc. Chem. Res.*, **39**, 568–575.

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.

Giasson,B.I. *et al.* (2001) A hydrophobic stretch of 12 amino acid residues in the middle of alpha-synuclein is essential for filament assembly. *J. Biol. Chem.*, **276**, 2380–2386.

Gromiha,M.M. *et al.* (2005) Application of residue distribution along the sequence for discriminating outer membrane proteins. *Comput. Biol. Chem.*, **29**, 135–142.

Gromiha,M.M. (2010) *Protein Bioinformatics: From Sequence to Function*. Elsevier/Academic Press, New Delhi.

Gromiha,M.M. and Suwa,M. (2006) Influence of amino acid properties for discriminating outer membrane proteins at better accuracy. *Biochim. Biophys. Acta*, **1764**, 1493–1497.

Iconomidou,V.A. *et al.* (2013) Identification of a novel 'aggregation-prone'/'amyloidogenic determinant' peptide in the sequence of the highly amyloidogenic human calcitonin. *FEBS Lett.*, **587**, 569–574.

Knowles,T.P.J. and Buehler,M.J. (2011) Nanomechanics of functional and pathological amyloid materials. *Nat. Nanotech.*, **6**, 469–479.

Kumar,S. *et al.* (2010) Identification and impact of aggregation prone regions in proteins and biotherapeutics. In: Wang,W. and Roberts,C.R. (eds) *Aggregation of Therapeutic Proteins*. John Wiley and Sons, Hoboken, NJ, pp. 103–118.

Liu,R. *et al.* (2012) Physicochemical strategies for inhibition of amyloid fibril formation: an overview of recent advances. *Curr. Med. Chem.*, **19**, 4157–4174.

Lopez de la Paz,M. and Serrano,L. (2004) Sequence determinants of amyloid fibril formation. *Proc. Natl Acad. Sci. USA*, **101**, 87–92.

MacPhee,C.E. and Dobson,C.M. (2000) Formation of mixed fibrils demonstrates the generic nature and potential utility of amyloid nanostructures. *J. Am. Chem. Soc.*, **122**, 12707–12713.

Maurer-Stroh,S. *et al.* (2010) Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat. Methods.*, **7**, 237–242.

Ou,Y.Y. *et al.* (2013) Classification of efflux proteins using efficient radial basis function networks with position-specific scoring matrices and biochemical properties. *Proteins*, **81**, 1634–1643.

Pawar,A.P. *et al.* (2005) Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases. *J. Mol. Biol.*, **350**, 379–392.

Pike,C.J. *et al.* (1995) Structure-activity analyses of beta-amyloid peptides: contributions of the beta 25-35 region to aggregation and neurotoxicity. *J. Neurochem.*, **64**, 253–265.

Roland,B.P. *et al.* (2013) A serendipitous survey of prediction algorithms for amyloidogenicity. *Biopolymers*, **100**, 780–789.

Rose,P.W. *et al.* (2013) The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Res.*, **41**, D475–D482.

Sawaya,M.R. *et al.* (2007) Atomic structures of amyloid cross-beta spines reveal varied steric zippers. *Nature*, **447**, 453–457.

Sonego,P. *et al.* (2008) ROC analysis: applications to the classification of biological sequences and 3D structures. *Brief. Bioinform.*, **9**, 198–209.

Tartaglia,G.G. *et al.* (2004) The role of aromaticity, exposed surface, and dipole moment in determining protein aggregation rates. *Protein Sci.*, **13**, 1939–1941.

Thangakani,A.M. *et al.* (2013) 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*, **14** (**Suppl. 8**), S6.

Thangakani,A.M. *et al.* (2012) How do thermophilic proteins resist aggregation? *Proteins*, **80**, 1003–1015.

Tjernberg,L. *et al.* (2002) Charge attraction and beta propensity are necessary for amyloid fibril formation from tetrapeptides. *J. Biol. Chem.*, **277**, 43243–43246.

Trovato,A. *et al.* (2007) The PASTA server for protein aggregation prediction. *Protein Eng. Des. Sel.*, **20**, 521–523.

Tsolis,A.C. *et al.* (2013) A consensus method for the prediction of 'aggregation-prone' peptides in globular proteins. *PLoS One*, **8**, e54175.

Witten,I.H. and Frank,E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd edn. Morgan Kaufmann, San Francisco, CA.