International Journal of Biological, Life and Agricultural Sciences
ISSN: 2415-6612
Vol:9, No:2, 2015

# Investigations of Protein Aggregation Using Sequence and Structure Based Features

M. Michael Gromiha, A. Mary Thangakani, Sandeep Kumar, D. Velmurugan

***Abstract*****—**The main cause of several neurodegenerative diseases such as Alzheimer, Parkinson and spongiform encephalopathies is formation of amyloid fibrils and plaques in proteins. We have analyzed different sets of proteins and peptides to understand the influence of sequence based features on protein aggregation process. The comparison of 373 pairs of homologous mesophilic and thermophilic proteins showed that aggregation prone regions (APRs) are present in both. But, the thermophilic protein monomers show greater ability to 'stow away' the APRs in their hydrophobic cores and protect them from solvent exposure. The comparison of amyloid forming and amorphous β-aggregating hexapeptides suggested distinct preferences for specific residues at the six positions as well as all possible combinations of nine residue pairs. The compositions of residues at different positions and residue pairs have been converted into energy potentials and utilized for distinguishing between amyloid forming and amorphous β-aggregating peptides. Our method could correctly identify the amyloid forming peptides at an accuracy of 95-100% in different datasets of peptides.

***Keywords*****—**Aggregation prone regions, amyloids, thermophilic proteins, amino acid residues, machine learning.

## I. Introduction

AGGREGATION is an ancient threat to productive protein folding and it is essential to overcome aggregation for the maintenance of metabolic flux and viability of cellular machineries. The aggregation of endogenous proteins causes several diseases in humans and animals. Aggregation is also a major hurdle in successful development of biopharmaceutical drug products [1]. Given the importance of aggregation in different areas of biology, it is important to elucidate different aggregation mechanisms and identify the probability of peptides to form amorphous β aggregates or amyloid fibrils. Several experimental studies have been carried out to understand the molecular determinants of aggregation, mutational effects as well as the influence of hydrophobic residues for promoting aggregation [2]-[6].

On the other hand, several computer approaches have been put forward to understand the mechanism of aggregation, aggregation strategies for mesophilic and thermophilic proteins and for predicting amyloid prone peptides [7]-[10]. These methods are mainly based on amino acid properties, such as hydrophobicity, charge distribution and propensity of β-strands [8]-[10]. Further, several structure-based models and empirical equations have been proposed to predict aggregation prone regions and change in aggregation propensity/rate due to mutation [11]-[17]. Agrawal et al. [1] and Belli et al. [18] have reviewed several commonly available aggregation prediction tools and discussed their advantages and shortcomings with applications.

In this work, we have identified the aggregation prone regions (APRs) in a set of 373 pairs of mesophilic and thermophilic proteins and analyzed them to understand the strategies evolved by the thermophiles to resist aggregation. Further, we have systematically analyzed the preferences of amino acid residues in amyloid fibril forming peptides (referred to as amyloids) and amorphous β-aggregating peptides (referred to as non-amyloids) at different positions and all possible nine residue pairs in experimentally known hexapeptides. The analysis showed the presence of several similarities and differences at different positions and residue pairs of amyloids and non-amyloids. These preferences have been converted into potentials and utilized them for discriminating amyloid-forming peptides from non-amyloids. The salient features of the results will be discussed.

Prof. Dr. Michael Gromiha Maria Siluvay is with IIT Madras, India (e-mail: gromiha@iitm.ac.in).

## II. APRs in Mesophilic and Thermophilic Proteins

### A. Dataset

We have used a dataset of 373 pairs of thermophilic and mesophilic proteins compiled by [19] in this study. The dataset has the following features: (i) multi-domain proteins were divided into separate single domains, (ii) a domain has no more than 400 residues, (iii) if one partner of the pair had longer sequences at the N or C termini, the extended segment of residues were truncated, (iv) the difference in the length between the proteins in a pair was no more than 10%, (v) number of residues that lack 3D coordinates were no more than 10% and (vi) the structural alignment score computed with Maxsub was greater than 70%.

### B. Aggregation Score in Mesophilic and Thermophilic Proteins

We have computed the aggregation score in mesophilic and thermophilic protein sequences using the programs, TANGO [12], PAGE [14] and Waltz [10] as described earlier [20]. TANGO is based on the physicochemical principles of β-sheet formation, extended by the assumption that the core regions of an aggregate are fully buried. PAGE is based on physicochemical properties and computational design of β-aggregating peptide sequences. Waltz uses position specific scoring matrices and performs well in recognizing polar APRs [10]. Recent study showed that the accuracy of currently available computational tools for prediction of aggregation prone regions in proteins are approximately 80% [21].

We have computed the aggregation score for all mesophilic and thermophilic proteins and their difference using TANGO. The results were grouped into three different clusters: (i) both mesophiles and thermophiles have similar aggregation scores, (ii) mesophiles with higher aggregation scores and (iii) high scores for thermophilic proteins. We observed that mesophilic proteins tend to have lower aggregation scores. Approximately, 55% of the mesophilic proteins have the aggregation score of less than 600. The higher aggregation scores (>800) are more frequent for the thermophilic proteins. This might be due to the differences in amino acid composition among the APRs in thermophilic and mesophilic proteins [7].

### *C. Aggregation Prone Regions Identified Using Specific Patterns*

Lopez de la Paz and Serrano have studied the link between amino acid sequence and amyloid fibril formation [22]. They have used a de novo designed amyloid hexa-peptide STVIIE and mutated each of the 6 positions with all the possible 19 natural amino acids and studied amyloid fibril formation. The authors have described two amyloidogenic sequence patterns stated below:

Pattern 1

{P}1-{PKRHW}2-[VLS(C)WFNQE]3-[ILTYWFNE]4-[FIY]5-{PKRH}6 for acidic pH.

Pattern 2

{P}1-{PKRHW}2-[VLS(C)WFNQ]3-[ILTYWFN]4-[FIY]5-{PKRH}6 for neutral pH.

These sequence patterns are written in PROSITE format. The numbers 1 to 6 represent positions in the hexa-peptide. The curly ({ }) and the straight ([ ]) brackets indicate disallowed and allowed residues at a given position.

A third sequence pattern has been described by Tjenberg and coworkers [23]:

Pattern 3

[KE]1-[FV]2-[FV]3-[EK]4 where the residue at position 1 is not the same as the one at position 4.

We have evaluated the existence of tetra-peptide and hexa-peptide amyloid-like fibril forming patterns in mesophilic and thermophilic proteins. These patterns were detected by using ScanProsite (http://prosite.expasy.org/scanprosite/) and http://www.bioinformatics.org/sms2/protein_pattern.html, a pattern matching tool. The tetra-peptide pattern (pattern 3) showed 18 hits in 17 mesophilic sequences whereas it showed 40 hits in 38 thermophilic sequences. Similarly, acidic pH hexa-peptide pattern (pattern 1) showed 811 and 894 hits in mesophilic and thermophilic proteins, respectively. The neutral pH hexa-peptide pattern (pattern 2) is a subset of acidic pH pattern (pattern 2). This pattern has 538 and 529 hits in the mesophilic and thermophilic proteins, respectively. We have also evaluated the overlap of these patterns with APRs detected by TANGO/Page and Waltz programs. Pattern 2 showed the greatest number of overlaps with APRs. Furthermore, the number of matches between patterns and predicted APRs was greater for thermophilic proteins than mesophilic proteins. Interestingly, the incidence of these patterns is very similar between thermophilic and mesophilic proteins.

We have also compared the existence of experimentally validated aggregating peptide sequences in mesophilic and thermophilic proteins by scanning their sequences against the library of 517 experimentally validated peptide sequences. We found that mesophilic and thermophilic proteins have 19 and 22 aggregating peptide sequences, respectively. These results indicate that the thermophilic proteins may also aggregate and form amyloid-like fibrils in a manner similar to their mesophilic homologues because sequence features that facilitate cross-β motif were observed in the thermophilic proteins as well.

Further, we have analyzed the variation of aggregation score with surrounding hydrophobicity and solvent accessibility [7]. We observed that the proteins with low aggregation score prefer to have lower surrounding hydrophobicity and are more exposed to solvent. The trend is opposite for the proteins with high aggregation scores. The correlations of aggregation score with hydrophobicity and solvent accessibility are slightly stronger for the thermophilic proteins. Hence, thermophilic proteins are able to 'stow away' their APRs in more hydrophobic regions than the mesophilic proteins.

## III. Sequence Analysis of Amyloid Forming Peptides and Non-Amyloids

### *A. Dataset*

We have developed a dataset, which contains 139 amyloid and 168 non-amyloid hexapeptides, which have been often used in experiments to grow amyloid-fibrils [10], [12]. These data have been collected from the careful search on the literature.

### *B. Amino Acid Composition of Amyloid Forming Peptides and Non-Amyloids at Different Positions*

The amino acid composition for the set of amyloids and non-amyloids at different positions has been computed using the number of amino acids of each type and the total number of residues in their respective positions. It is defined as [24], [25]:

$$\mathrm{Comp}(i,j) = \Sigma\, n_{ij}/N_j \qquad (2)$$

where i and j stands for the 20 amino acid residues and six positions, $n_{ij}$ is the number of residues of each type i at position j and N is the total number of residues at position j.

The amino acid compositions at six positions of amyloid forming hexapeptides showed that specific residues are preferred at some positions [26]. Especially, Glu prefers to accommodate at position 6 compared with other positions as well as other amino acid residues; Ile is dominant in positions 4 and 5 and the difference is highly significant; position 1 is

accommodated by Ser; Thr and Val showed their preference at positions 2 and 3, respectively [26]. In amorphous peptides, Ser prefers position 1; Ala and Thr show high preference at position 2; Val in position 3; Ile and Leu in position 4; Phe and Ile in position 5, and Gln in position 6. Although some of the features are similar to amyloids and non-amyloids the occurrence of several amino acids are different from each other. For example, Ala in position 2, Asn in position 4, Gly in position 6 and so on [26]. These differences are helpful for discriminating amyloids and non-amyloids.

### *C. Preference of Residue Pairs in Amyloid and Amorphous Peptides*

We have computed residue pair compositions in amyloid and amorphous hexa-peptides and the results revealed the preference of distinct residue pairs in amyloids and amorphous peptides [27]. For example, the most preferred alternate residue pairs are Gly-Thr, at positions 1-2; Thr-Val at 2-3; Phe-Phe, at 3-4; Trp-Ile at 4-5, Ile-Glu at 5-6, Ser-Val at 1-3; Gln-Ile at 2-4, Ser-Phe at 3-5 and Ile-Glu at 4-6. The preferred residue pairs in amorphous peptides are Lys-Ala at positions 1-2, Met-Phe at 2-3, Phe-Phe at 3-4, Ile-Ile at 4-5, Ile-Ser at 5-6, Ser-Val at 1-3, Thr-Ile at 2-4, Phe-Ile at 3-5 and Ile-Glu at 4-6. These residue pairs along with other preferred pairs can be used for distinguishing between them.

## IV. Discrimination of Amyloid Forming Peptides and Non-Amyloids

### *A. Energy Potentials*

We have converted the composition of amino acid residues at different positions of hexa-peptides (Eqn. 2) into propensities by normalizing the composition with overall composition of globular proteins [24], [25]. The propensity of amino acid residues at different positions is given by

$$\text{Propen}(i,j) = \text{Comp}(i,j)/\text{Compglob}(i) \quad (3)$$

where, Compglob(i) is the composition of residue i obtained with a set of globular proteins [24], [25].

These amino acid propensities at each of position of amyloid and non-amyloid peptides were treated as partition functions and converted into thermodynamic energy potential by using:

$$\phi(i,j) = -RT \ln \text{Propen}(i,j) \quad (4)$$

where, i and j are the 20 amino acid residues and six positions respectively. We have derived the energy potentials for both amyloid and amorphous peptides, which can be used to distinguish these types of peptides [28].

We have followed a similar procedure to derive the potentials for all the nine residue pairs (20x20 matrices). The specific residue pairs, which showed significant differences in energy between amyloid and amorphous peptides are given below: KC, MC, MH, MY, NH and WT at positions 1-2; CW, EC, FW, MF, QF and VF at positions 2-3, CL, CY, FW, VN, VW, WC and WW at positions 3-4, FF, IF, WI and WY at positions 4-5, FW and WC at positions 5-6, EC, HC, HW, KW, MF, MV, MW, TW, VW and WV at positions 1-3, CC, CN, FW, HC, LW, MF, QI, VW and YW at positions 2-4, QY, VF and WY at position 3-5 and FC, NF and WC at positions 4-6.

### *B. Machine Learning Techniques and Assessment of Predictive Ability*

We have analyzed several machine learning techniques implemented in WEKA program [29] for discriminating between amyloid and non-amyloid peptides. WEKA includes several methods based on different machine learning techniques such as Bayesian function, Neural network, Radial basis function network, Logistic function, Support vector machine, Regression analysis, Nearest neighbor, Meta learning, Decision tree and Rules. The details of all these methods are available in our earlier articles [30]. We have used the energy potentials and selected amino acid properties as input features for the methods.

We have performed 20-fold, 10-fold and 5-fold cross-validation tests for assessing the validity of the present work. In this method, the data set is divided into n groups, n-1 of them, are used for training and the rest is used for testing the method. The same procedure is repeated for n times so that each data is used at least once in the test.

We have used different measures, such as sensitivity, specificity and accuracy, to assess the performance of machine learning methods towards discriminating between amyloid and non-amyloid peptides. The term sensitivity shows the correct prediction of amyloid peptides, specificity is the correct prediction of non-amyloid peptides and accuracy indicates the overall assessment. These terms are defined as:

$$\text{Sensitivity} = TP/(TP+FN)$$
$$\text{Specificity} = TN/(TN+FP)$$
$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN),$$

where, TP, FP, TN and FN refer to the number of true positives, false positives, true negatives and false negatives, respectively.

### *C. Discrimination of Amyloid Forming Peptides and Non-Amyloids*

We have utilized several machine learning techniques for discriminating between amyloid and non-amyloid peptides as described in the Methods section. Overall, most algorithms showed similar performance. In a 10-fold cross-validation method, we obtained an accuracy of 82.1% using statistically derived position-specific energy potentials. The sensitivity and specificity are 79.9% and 83.9%, respectively. Combining these energy potentials with three amino acid properties, hydrophobicity, isoelectric point and long-range non-bonded energy improved the accuracy marginally to 82.7% (sensitivity, 81.3%; specificity, 83.9%). The method was also tested with 5-fold and 20-fold cross-validations and the accuracies are 80% and 81.1%, respectively.

Further, we have utilized the energy potentials derived with residue pairs, which showed the accuracy of 93.7% in a set of 179 amyloid and 168 amorphous peptides. The sensitivity and specificity are 95.5% and 91.7%, respectively. The method, GAP was examined with 15 amyloid peptides from the Protein Data Bank, which showed the sensitivity of 93.3% [27].

GAP was tested on 310 amyloid-fibril forming peptides of different lengths. For the peptides of more than six residues, we divided the peptide sequence into six residue long windows that slide by one residue at a time. For each window, we have computed the scores for amyloid-fibril formation and amorphous aggregation. These scores are summed over all the windows to obtain total scores for amyloid fibril formation and amorphous aggregation for the whole peptide. The better 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 Fig. 1. The accuracy is 100% for most of the peptides. Specifically, all 14-20 residues long peptides and those longer than 22 residues are correctly predicted to be amyloid fibril forming peptides. For most of the remaining peptides, the accuracy is more than 90%. In all these cases, GAP missed only one or two peptides for a given length. Overall, 302 out of 310 (97.4%) peptides in Amyl310 dataset are correctly predicted to be amyloid-fibril forming [27].

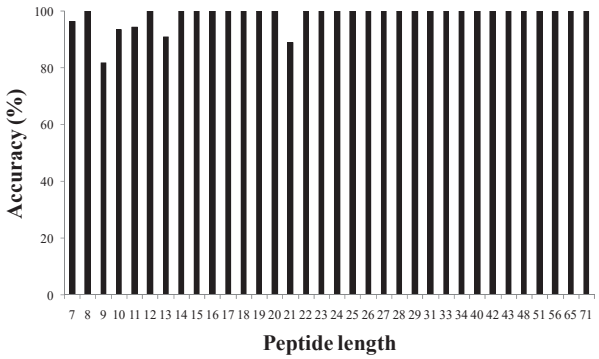


Fig. 1 Prediction performance of GAP on amyloid-fibril forming peptides of different lengths

GAP was also tested on 48 experimentally determined amyloid fibril forming peptide segments, of different lengths, from 33 well-known amyloidogenic proteins [31]. It correctly predicted 47 (98%) of them. Further, developers of WALTZ [10] had bench-marked the performance of their method by using twelve sup35-derived 10-resudues long peptides that were shown to form amyloid-fibrils experimentally. The performance of GAP along with other prediction methods TANGO, WALTZ, and Amylpred2 is shown in Table I. GAP correctly predicted all the 12 peptides (100% sensitivity).

TABLE I
PERFORMANCE OF DIFFERENT PREDICTION METHODS ON SEQUENCES OF SUP35-DERIVED AMYLOID-FIBRIL FORMING PEPTIDES

| Sequence | WALTZ | TANGO | Amylpred2 | GAP |
|---|---|---|---|---|
| GNNQQNYQQY | + | - | -- | + |
| YSQNGNQQQG | - | - | - | + |
| RYQGYQAYNA | + | - | - | + |
| GGYYQNYQGY | + | - | - | + |
| YQNYQGYSGY | + | - | - | + |
| YSGYQQGGYQ | - | - | - | + |
| YQQGGYQQYN | + | - | - | + |
| PQGGRGNYKN | - | - | - | + |
| NFNYNNNLQG | + | - | - | + |
| YNNNLQGYQA | + | - | - | + |
| NLQGYQAGFQ | + | - | - | + |
| Total 12 | 8 | 0 | 0 | 12 |
| Sensitivity | 66.6% | 0% | 0% | 100% |

+ or- represent correct or incorrect predictions, respectively. The sequences were taken from [10].

## ACKNOWLEDGMENT

AMT and DV thank the Bioinformatics Facility of University of Madras for computational facilities. MMG wishes to thank IIT Madras for infrastructure facilities.

## REFERENCES

[1] Agrawal NJ, Kumar S, Wang X, Helk B, Singh SK, Trout BL 2011. Aggregation in protein-based biotherapeutics: Computational studies and tools to identify aggregation-prone regions. J Pharm Sci 2011, 100:5081-5095.
[2] Wurth C, Guimard NK, Hecht MH (2002) Mutations that reduce aggregation of the Alzheimer's Aβ42 peptide: an unbiased search for the sequence determinants of Aβ amyloidogenesis. J Mol Biol 319: 1279–1290
[3] de Groot NS, Aviles FX, Vendrell J, Ventura S (2006) Mutagenesis of the central hydrophobic cluster in Aβ42 Alzheimer's peptide. Side-chain properties correlate with aggregation propensities. FEBS J 273: 658–668
[4] Kim W, Hecht MH (2006) Generic hydrophobic residues are sufficient to promote aggregation of the Alzheimer's Aβ42 peptide. Proc Natl Acad Sci USA 103: 15824–15829
[5] Luheshi LM et al (2007) Systematic in vivo analysis of the intrinsic determinants of amyloid β pathogenicity. PLoS Biol 5: e290
[6] Winkelmann J, Calloni G, Campioni S, Mannini B, Taddei N, Chiti F (2010) Low-level expression of a folding-incompetent protein in Escherichia coli: search for the molecular determinants of protein aggregation in vivo. J Mol Biol 398: 600–613
[7] Thangakani AM, Kumar S, Velmurugan D, Gromiha MM. How do thermophilic proteins resist aggregation? Proteins. 2012;80:1003-15.
[8] Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nat Biotechnol 22: 1302–1306
[9] Conchillo-Solé O, de Groot NS, Avilés FX, Vendrell J, Daura X, Ventura S (2007) AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. BMC Bioinformatics 8: 65
[10] Maurer-Stroh S et al (2010) Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. Nat Methods 7: 237–242.
[11] DuBay KF, Pawar AP, Chiti F, Zurdo J, Dobson CM, Vendruscolo M. Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains. J Mol Biol 2004, 341: 1317–1326.
[12] Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nat Biotechnol 2004, 22: 1302–1306.

[13] Pawar AP, Dubay KF, Zurdo J, Chiti F, Vendruscolo M, Dobson CM. Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases. J. Mol. Biol. 2005, 350:379-392.
[14] Tartaglia GG, Cavalli A, Pellarin R, Caflisch A. Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. Protein Science 2005, 14:2723-2734.
[15] Galzitskaya OV, Garbuzynskiy SO, Lobanov MY. Prediction of amyloidogenic and disordered regions in protein chains. PLoS Comput Biol 2006, 2: e177.
[16] Thompson MJ, Sievers SA, Karanicolas J, Ivanova MI, Baker D, Eisenberg D. The 3D profile method for identifying fibril-forming segments of proteins. Proc Natl Acad Sci USA 2006, 103: 4074–4078.
[17] Zibaee S, Makin OS, Goedert M, Serpell LC. A simple algorithm locates β-strands in the amyloid fibril core of α-synuclein, Aβ, and tau using the amino acid sequence alone. Protein Sci 2007, 16: 906–918.
[18] Belli M, Ramazzotti M, Chiti F. Prediction of amyloid aggregation in vivo. EMBO Rep. 2011;12:657-63.
[19] Glyakina AV, Garbuzynskiy SO, Lobanov MY, Galzitskaya OV. (2007) Different packing of external residues can explain differences in the thermostability of proteins from thermophilic and mesophilic organisms. Bioinformatics. 23(17):2231-8.
[20] Wang X, Singh SK, Kumar S. (2010) Potential Aggregation-Prone Regions in Complementarity-Determining Regions of Antibodies and Their Contribution Towards Antigen Recognition: A Computational Analysis. Pharm Res (2010) 27:1512–1529.
[21] Kumar S, Wang X, Singh SK. (2010) Identification and impact of aggregation prone regions in proteins and therapeutic mAbs. In: Wang and W, Roberts C, editors. Aggregation of therapeutic proteins. US: Wiley; Hoboken, 2010; pp 103 – 118.
[22] Lopez de la Paz, M, Serrano. (2004) Sequence determinants of amyloid fibril formation. Proc Natl Acad Sci USA, 101, 87-92.
[23] Tjernberg, L., Hosia, W., Bark, N., Thyberg, J. Johansson, J. Charge attraction and beta propensity are necessary for amyloid fibril formation from tetrapeptides. J Biol Chem 2002, 277, 43243-6.
[24] Gromiha MM, Suwa M. A simple statistical method for discriminating outer membrane proteins with better accuracy. Bioinformatics. 2005;21:961-8
[25] Gromiha MM. (2010) Protein Bioinformatics: From Sequence to Function, Elsevier/Academic Press.
[26] Gromiha MM, Thangakani AM, Kumar S, Velmurugan D. (2012) Sequence analysis and discrimination of amyloid and non-amyloid peptides. Comm. Comp. Inf. Sci. 304, 447-452 (2012).
[27] Thangakani AM, Kumar S, Nagarajan R, Velmurugan D, Gromiha MM. GAP: towards almost 100 percent prediction for β-strand-mediated aggregating peptides with distinct morphologies. Bioinformatics. 2014;30(14):1983-90.
[28] Thangakani AM, Kumar S, Velmurugan D, Gromiha MM. 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 Suppl 8:S6
[29] Witten IH, Frank E: Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.
[30] Gromiha MM, Suwa M. Influence of amino acid properties for discriminating outer membrane proteins at better accuracy. Biochim Biophys Acta. 2006;1764:1493-7.
[31] Tsolis AC1, Papandreou NC, Iconomidou VA, Hamodrakas SJ. A consensus method for the prediction of 'aggregation-prone' peptides in globular proteins. PLoS One. 2013;8(1):e54175