JPP JOURNAL OF PROTEINS AND PROTEOMICS
7(4), 2016, pp. 259-268

Research Article

ORIGIN OF PROTEIN AGGREGATION: IDENTIFICATION OF SOME CHARACTERISTIC TRAITS IN STRUCTURED AND INTRINSICALLY DISORDERED PROTEINS

Uttam Pal, Anupam Roy, Supriya Das, Swagata Das, Mangaldeep Kundu, Khyati Bagga and Nakul Chandra Maiti*

Structural Biology and Bioinformatics Division, Council of Scientific and Industrial Research (CSIR)-Indian Institute of Chemical Biology (IICB), 4, Raja S.C. Mullick Road, Kolkata 700032, India

Abstract: Proteins containing amyloidogenic regions are prone to form aggregates and undergo amyloidosis. Literature shows that both the structured and disordered proteins might be amyloidogenic. However, the amyloidogenic proteins differ from their non-amyloidogenic counterpart in various physicochemical properties and, therefore, based on these properties they can be categorized. Our analysis also indicated that the sequence composition and the physicochemical properties such as isoelectric point (pI), hydrophobicity, aliphatic index (AI) and instability index (II) of amyloidogenic and non-amylidogenic proteins differed in large extent. However, unique to our finding is that such differences appeared to be exclusive for intrinsically disordered proteins. Structured amyloidogenic proteins were found to be more similar to their non-amyloidogenic counterparts except for their sequence composition and the distribution of their isoelectric points. In this report, we have shown from the sequence analysis that the distinction between amyloidogenic and non-amyloidogenic proteins is much wider in case of disordered proteins than the structured proteins. Our finding also suggests why structured proteins that are amyloidogenic require drastic change in the solution condition to undergo amyloid formation. The results have implication in developing better algorithms for the detection and differentiation of amyloidogenic proteins.

***Keywords*:** Amyloidogenic; Disordered protein; Physicochemical properties; Sequence composition; Aggregation

Note : Coloured Figures and Supplementary Information available on Journal Website in "Archives" Section

Introduction

In medicine, amyloidosis defines the diseases that result from the abnormal deposition of particular proteinaceous masses, called the amlyloids, in various tissues and organs of the body (Westermark *et al.*, 2005). Some of the proteins that are linked to amyloid formation are intrinsically disordered or lack any globular structure (Xie *et al.*, 2007a, 2007b). Natively disordered protein α-synuclein is a part of the abnormal protein aggregate found in Lewy bodies. Similarly, tau, amyloid-β and many other proteins are natively unfolded and associated with amyloid diseases (Taylor *et al.*, 2002). However, many globular proteins that have compact fold with defined secondary and tertiary structures are also linked to amyloid formation and diseases progression (Chiti and Dobson, 2009). Examples of such proteins include β2-microglobulin, myoglobin, lysozyme, insulin and many others. Moreover, natively unfolded or folded proteins, which are not linked to any disease are also seen to form amyloid fiber under

Corresponding Author: **Nakul C. Maiti**
E-mail: ncmaiti@iicb.res.in

Received: October 31, 2016
Accepted: November 26, 2016
Published: December 14, 2016

suitable solution conditions such as low pH, high ionic strength and in the presence of different co-solvents (Das *et al.*, 2013; Fändrich *et al.*, 2001; McParland *et al.*, 2000; Nielsen *et al.*, 2001; Vernaglia *et al.*, 2004). Apart from that, the fibril formation propensity of a protein largely depends on its sequence composition (Bemporad *et al.*, 2006; Calamai *et al.*, 2003; Chiti *et al.*, 2003). For example, mutations in the sequence can prevent protein fibril formation, especially if the mutation is a beta-sheet breaker, such as proline residue in a certain position (Chaitanya *et al.*, 2013; Morimoto *et al.*, 2002). Mutations in a protein sequence that destabilize the secondary structure, sometimes promote aggregation (Meuvis *et al.*, 2010; Morimoto *et al.*, 2002). Disordered proteins seldom contain a simple amino acid sequence with compositional bias, known as low complexity regions, which usually enhances the flexibility of unfolded protein (Das *et al.*, 2014; Dosztányi *et al.*, 2006; Wootton, 1994; Wootton and Federhen, 1993). On the other hand, structured proteins contain more hydrophobic amino acids and often require external stimuli to partially unfold and further misfold to form beta-sheet rich amyloid aggregates (Chiti *et al.*, 2002). Therefore, among the many other factors, the composition and the content of amino acid residues in a protein sequence, the stability of a protein in solution and the hydrophobicity have become pivotal to realize the fibril formation and associated phenomena of amyloidosis.

In this investigation, we used *in silico* methods to characterize the amyloidogenic and non-amyloidogenic structured and disordered human proteins. Disordered and structured proteins were obtained from DisProt (Sickmeier *et al.*, 2007) and Protein Data Bank (PDB), respectively. DisProt enlists proteins that lack fixed three dimensional structure in their putatively native states either in their entirety or in part. We defined structured proteins as amyloidogenic if it contained amyloidogenic regions in its sequence. In addition to sequence composition, various physicochemical properties such as isoelectric point, hydrophobicity in terms of grand average hydropathy (GRAVY), aliphatic index and instability index were theoretically determined and statistically analyzed to segregate the different groups of proteins.

Methods

Building the dataset

The dataset, used in our analysis contains sequence and other related information such as DistProt and UniProt accession codes of a total of 481 human proteins. Information about 217 disordered human proteins was retrieved from DisProt (Sickmeier *et al.*, 2007) database (Release 6.02, May 2013). Both the experimentally proven and theoretically predicted proteins from DisPort were included in the dataset. Information about 264 human proteins with unique UniProt ID, sharing less than 30% sequence identity and with reported high resolution (more than 1.5Å) crystallographic structure was obtained from PDB (May 2013).

Determination of amyloidogenicity

Proteins containing amyloidogenic regions (ARs) are prone to aggregate and thus, amyloidogenic in nature. ARs in a protein sequence were identified by a web based computational tool, Waltz (Maurer-Stroh *et al.*, 2010). It is relatively a new prediction algorithm and recognizes amyloidogenic region for a given amino acid sequence. This method uses a position specific scoring matrix (PSSM) to determine the amyloid-forming segments. The PSSM matrix was derived from the "position specific" sequence composition and physical properties in a set of experimentally validated hexapeptide library. The algorithm does not consider the sequence composition or physicochemical properties of the peptide/protein as a whole. There are many other algorithms (Chennamsetty *et al.*, 2010; Fang and Fang, 2013; Garbuzynskiy *et al.*, 2010) for the detection of amyloidogenic protein and a consensus prediction is generally favored (Belli *et al.*, 2011). However, the noise in the detection of amyloidogenic proteins due to computational error was ignored based on the statistical ground that two different populations of protein (structured and intrinsically disordered) were analyzed and compared. Computational error being random in nature contributed equally to both the population analysis. Although it increased the noise, it did not bias the results.

Calculation of pI, GRAVY, AI, II and amino acid composition

Isoelctric point (pI), aliphatic index (AI) and instability index (II) of 217 unfolded proteins from DisProt and 264 structured proteins from PDB were calculated using ProtParam tool of ExPASy Proteomic server (Gasteiger *et al.*, 2005). Overall protein hydrophobicity was also calculated using ProtParam tool in terms of GRAVY, a measure of per residue hydrophobicity.

Statistical analysis

All the statistical analysis was performed in Wolfram Mathematica 9. Cramér-von Mises test (Anderson and Darling, 1952) was used to confirm whether the data was normally distributed or not. For the normally distributed data, mean, standard deviation (SD) and standard error of mean (SEM) were calculated. Significance of the mean differences was established with Student's t-test and the null hypotheses were rejected at 5 percent level of significance. Probability values of less than 0.0005 were considered highly significant and denoted by *** in the graphs. Likewise, the probability values in between 0.0005 and 0.005 were considered very significant and denoted by ** in the graphs and the rest were denoted by a single star; this convention was followed or otherwise mentioned. Normal distributions with mean and SD were fitted to the data and the distribution fit tests were performed to verify its consistency with the data. For the bimodal distributions, mixture of distributions was fitted to the data.

Results and Discussion

In the last decade, several sequence-based computational methods for the prediction of protein amyloidogenicity have been described (Chennamsetty *et al.*, 2010; Fang and Fang, 2013; Garbuzynskiy *et al.*, 2010; Maurer-Stroh *et al.*, 2010; Oldfield *et al.*, 2005). Contribution of different physicochemical properties to aggregation propensity has also been studied extensively by our group (Das *et al.*, 2013, 2014; Maity and Maiti, 2012; Pal *et al.*, 2016) and others (Bemporad *et al.*, 2006; Calamai *et al.*, 2003; Tcherkasskaya *et al.*, 2003). However, their relative contributions in different groups of protein were not well documented. Here, our analysis revealed that in structured and disordered group of proteins determinants of amyloidogenicity differs significantly.

Sequence analysis showed that 187 out of 217 (86%) disordered proteins and 217 out of 264 (82%) structured proteins in our dataset were amyloidogenic. The ARs constituted an average of 12% and 9% of the total length in disordered and structured group of amyloidogenic proteins, respectively. The computed physicochemical parameters namely pI, GRAVY, aliphatic index (AI) and instability index (II) of all the proteins are listed in Table S1 and Table S2. Mean values of all these parameters for the four groups of proteins are listed in the Table 1.

Statistical analysis showed that theoretical pI values followed a bimodal distribution for both the amyloidogenic and non-amyloidogenic structured/disordered proteins (Figure 1). pIs

Table 1
Physicochemical parameters of the amyloidogenic and non-amyloidogenic disordered and structured human proteins. The mean values, listed here were obtained from the fitted normal distributions

The values are represented as Mean±SEM. Where the parameters of amyloidogenic proteins differ significantly from non-amyloidogenic proteins are marked with asterisks. P-values are given in the supporting information.

Parameters	*Disordered proteins*		*Structured proteins*	
	Amyloidogenic	*Non-amyloidogenic*	*Amyloidogenic*	*Non-amyloidogenic*
pI (acidic)	5.66±0.06	5.53±0.24	5.86±0.06	5.71±0.14
pI (basic)	8.84±0.10***	9.80±0.26	8.39±0.09***	8.97±0.11
GRAVY	-0.50±0.02***	-0.91±0.08	-0.40±0.02	-0.45±0.05
AI	75.43±1.12**	60.75±4.03	79.89±0.98	78.20±1.17
II	49.10±0.99**	61.13±3.72	39.08±0.90	42.25±2.21

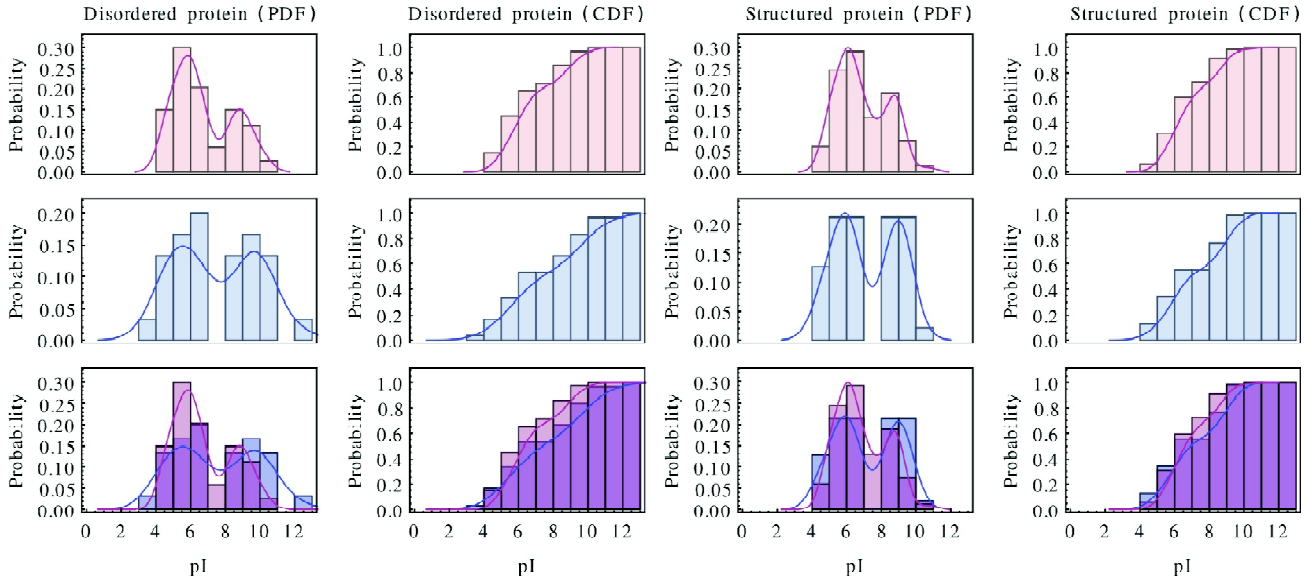

Figure 1: **Distribution of isoelectric points of amyloidogenic (light red) and non-amyloidogenic (light blue) disordered (column 1 and 2) and structured (column 3 and 4) human proteins.**
Histograms with the fitted distributions are shown. Both the Probability Density Function (column 1 and 3) and the Cumulative Density Function (column 2 and 4) are depicted. Last row is the overlap of the first two comparing the differences or the similarities. Horizontal and vertical axes represents the pI and probability or cumulative probability, respectively. pI followed a symmetric bimodal distribution for structured proteins whereas an asymmetric bimodal distribution for the disordered proteins. Mean pI of amyloidogenic proteins in the basic range were significantly lower than that of the non-amyloidogenic proteins.

were mostly distributed either in acidic or in basic regions but rarely at the neutral pH. On both sides they followed a normal distribution. Such multimodal distribution of pIs for the whole proteome is known in the literature (Kiraga *et al.*, 2007; Wu *et al.*, 2006). Moreover, the importance of the net charge of a polypeptide in determining its aggregation propensity has been long recognized (Calamai *et al.*, 2003; Chiti *et al.*, 2003; Zbilut *et al.*, 2004). In our analysis, however, comparison of the mean acidic pIs of the amyloidgenic and non-amyloidogenic proteins showed marginal difference. However, the mean pI of amyloidogenic proteins in basic range was significantly lower (p-values < 0.0005) than that of the non-amyloidogenic proteins. One more interesting observation was that, pIs of non amyloidogenic proteins were symmetrically distributed on both the acidic and basic domains. On the contrary, the distribution of pIs of amyloidogenic proteins was asymmetric and biased toward the acidic range. The corollary is that at neutral pH more amyloidogenic proteins are negatively charged when both the positive and negatively charged non-amyloidogenic proteins become equally abundant. As the pH value of a protein approaches its pI, charge and repulsive interactions decreases, and the hydrophobic interactions may lead to favorable association or amyloidogenesis. Thus, pI is a very important factor governing aggregation property of a protein. Buffering near to the pI of a protein could change the course of the aggregation and amyloid formation process (Idicula-Thomas and Balaji, 2005). Low pH is often found to cause partial or large unfolding of the local structure that leads to protein aggregation and fiber formation. Amyloidogenic proteins populated the lower pI range and thus, indicated that pH of the medium might have substantial effect on the protein stability and aggregation propensity.

Hydrophobicity is regarded as a major contributing factor in aggregation and folding (Calamai *et al.*, 2003; Monsellier *et al.*, 2008; Routledge *et al.*, 2009; Zbilut *et al.*, 2004). Consequently, amyloidogenic and non-amyloidogenic disordered proteins showed very significant (p-value < 0.0005) variation in GRAVY (Figure 2). However, GRAVY distribution of amyloidogenic and non-amyloidogenic structured proteins did not differ significantly (Figure 2). The GRAVY distribution was

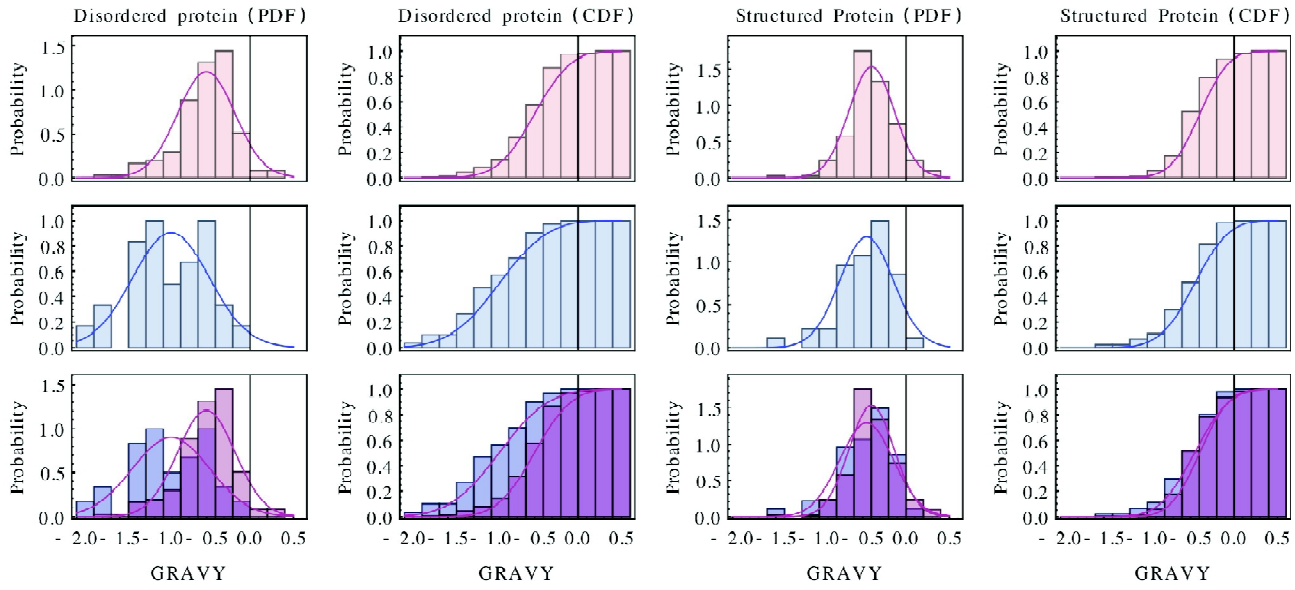

Figure 2: **Distribution of GRAVY of amyloidogenic (light red) and non-amyloidogenic (light blue) disordered (column 1 and 2) and structured (column 3 and 4) human proteins.**
Histograms with the fitted distributions are shown. Both the Probability Density Function (column 1 and 3) and the Cumulative Density Function (column 2 and 4) are depicted. Last row is the overlap of the first two comparing the differences or the similarities. Horizontal and vertical axes represents the GRAVY and probability or cumulative probability, respectively. Mean GRAVY of the disordered amyloidogenic proteins was significantly higher than that of the disordered non-amyloidogenic proteins and was comparable to the structured proteins.

predominantly in the hydrophilic regions, however, amyloidogenic disordered proteins showed significantly less hydrophilicity than their non-amyloidogenic counterpart. Interestingly, GRAVY distribution of disordered amyloidogenic proteins was observed to be comparable with that of the structured proteins, reasserting that a well organized hydrophobic core is crucial for folded structure but exposed hydrophobic residues for amyloid formation.

The amyloidogenic disordered proteins were found to have significantly higher aliphatic index (AI) (p-value < 0.005) than the non-amyloidogenic proteins of the same group (Figure 3). The AI of proteins of thermophilic bacteria was used as a measure of proteins thermostability (Ikai, 1980). Thermal stability of proteins provides insight into the disordered proteome (Galea *et al.*, 2009). The higher index values indicated larger thermostability of globular proteins. However, the mean AIs for structured proteins did not differ much among the amyloidogenic and non-amyloidogenic groups (Figure 3). When we compared the thermostability of the disordered proteins with the structured proteins, it was found that the thermostability of the amyloidogenic disordered proteins was similar to that of the structured proteins, whereas the non-amyloidogenic disordered proteins had much less thermal stability.

Instability index (II) (Guruprasad *et al.*, 1990) echoed that the structured proteins were more stable than the disordered proteins (Figure 4). However, within the disordered domain II of the amyloidogenic and non-amyloidogenic groups differed very significantly (p-value < 0.005). Amyloidogenic diordered proteins were found to be more stable and the II was comparable to that of the structured proteins. The non-amyloidogenic disordered proteins showed higher II and thus indicated less stability than the amyloidogenic disordered proteins or the structured proteins.

Amyloidogenic and non-amyloidogenic structured proteins, therefore, could not be discriminated based on the physicochemical properties such as, GRAVY, AI or II except for their pI, which is most affected by the solution condition. The results add another perspective to the theory that structured proteins, be it amyloidogenic or non-amyloidogenic, are inherently stable and requires drastic change in

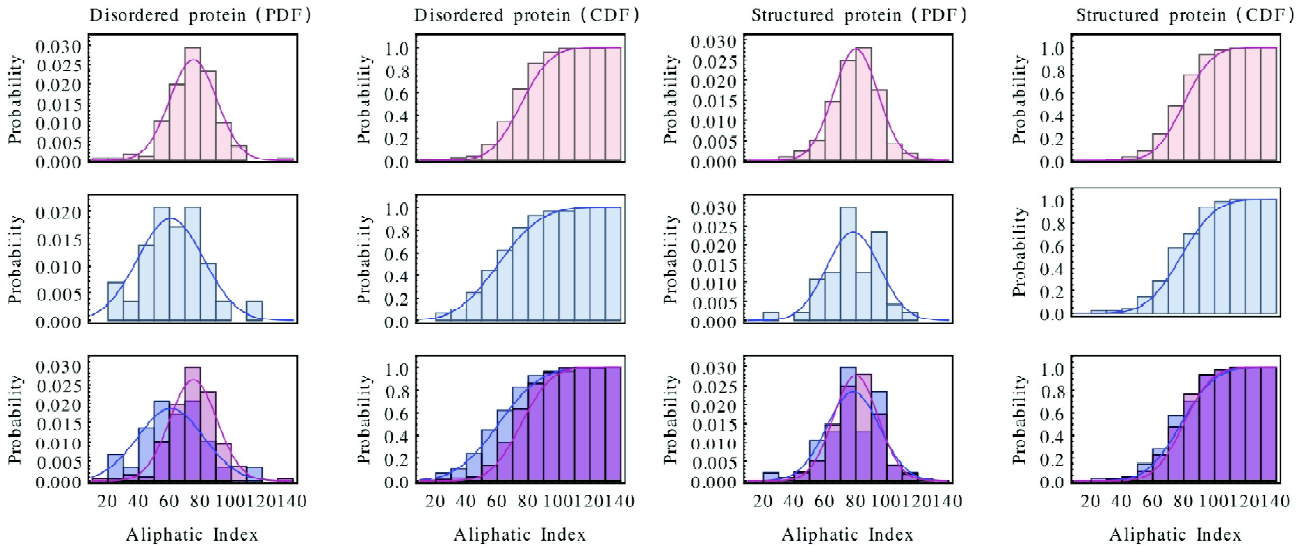

Figure 3: **Distribution of aliphatic index of amyloidogenic (light red) and non-amyloidogenic (light blue) disordered (column 1 and 2) and structured (column 3 and 4) human proteins.**
Histograms with the fitted distributions are shown. Both the Probability Density Function (column 1 and 3) and the Cumulative Density Function (column 2 and 4) are depicted. Last row is the overlap of the first two comparing the differences or the similarities. Horizontal and vertical axes represents the AI and probability or cumulative probability, respectively. Mean AI of the disordered amyloidogenic proteins was significantly higher than that of the disordered non-amyloidogenic proteins and was comparable to the structured proteins.

the solution conditions to undergo fibrillation or amyloidogenesis (Fändrich *et al.*, 2001; McParland *et al.*, 2000; Nielsen *et al.*, 2001; Vernaglia *et al.*, 2004). On the contrary, all the physicochemical properties of disordered amyloidogenic proteins differed significantly from their non-amyloidogenic counterpart. One more striking feature in these distributions was that the spread of the distributions for the disordered amyloidogenic proteins were much narrower, thus, indicating towards a stringent selection rule. Therefore, the disorder prediction prior to amyloidogenicity calculation or a combination of the two could provide a much better detection tool of aggregation propensity.

The differences in amino acid compositions between structured and disordered proteins (Ferron *et al.*, 2006; Kovačević, 2012) or amyloidogenic and non-amyloidogenic proteins (de Groot *et al.*, 2006; Pawar *et al.*, 2005) are yet another extensively studied factors. Using this difference it was possible to create some programs for prediction of disordered and amyloidogenic regions. Disordered and structured human proteins in our database also showed significant differences in their amino acid composition. It was found that apart from Lys, Asp, Thr, Asn, Gly, Leu and Met the distribution of other amino acid residues greatly differed between the two groups (Figure S1). However, Our main focus was to compare the sequence composition of amyloidogenic and non-amyloidogenic proteins in these two categories (Figure 5). The amino acids those were found to differ between amyloidogenic and non-amyloidogenic proteins in the structured group, also differed significantly in the disorder group. In both the structured and disordered groups, the content of Asn, Ile, Phe and Tyr was significantly higher among the amyloidogenic proteins. Except Asn, all are hydrophobic. Asn has an uncharged polar side chain. However, contribution of these hydrophobic amino acids could not discriminate the GRAVY of structured amyloidogenic and non-amyloidogenic proteins. More variations in the amino acid composition of amyloidogenic and non-amyloidogenic proteins were observed in the disordered group. In this category, amyloidogenic proteins were found to have more hydrophobic amino acids (Leu and Met) and significantly less charged amino acids (Arg, Glu and Lys); Pro was also found to differ. Abundance in hydrophobic amino acids along with the less number of charged amino acids contributed to the greater

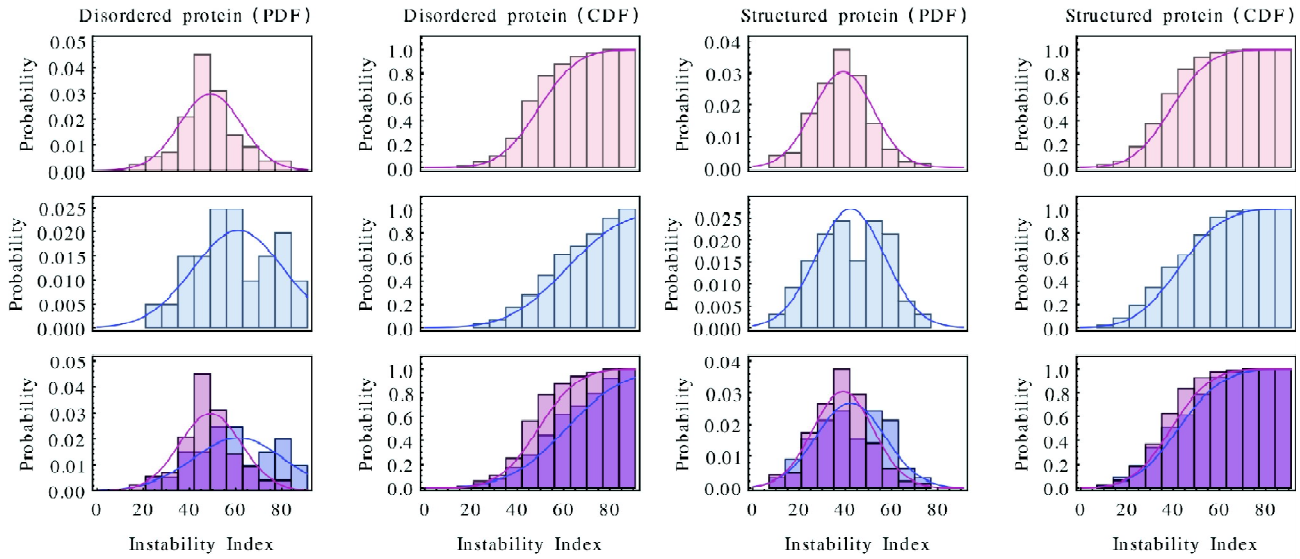

Figure 4: **Distribution of instability index of amyloidogenic (light red) and non-amyloidogenic (light blue) disordered (column 1 and 2) and structured (column 3 and 4) human proteins.**
Histograms with the fitted distributions are shown. Both the Probability Density Function (column 1 and 3) and the Cumulative Density Function (column 2 and 4) are depicted. Last row is the overlap of the first two comparing the differences or the similarities. Horizontal and vertical axes represents the II and probability or cumulative probability, respectively. Mean II of the disordered amyloidogenic proteins are significantly lower than that of the disordered non-amyloidogenic proteins and was comparable to the structured proteins.

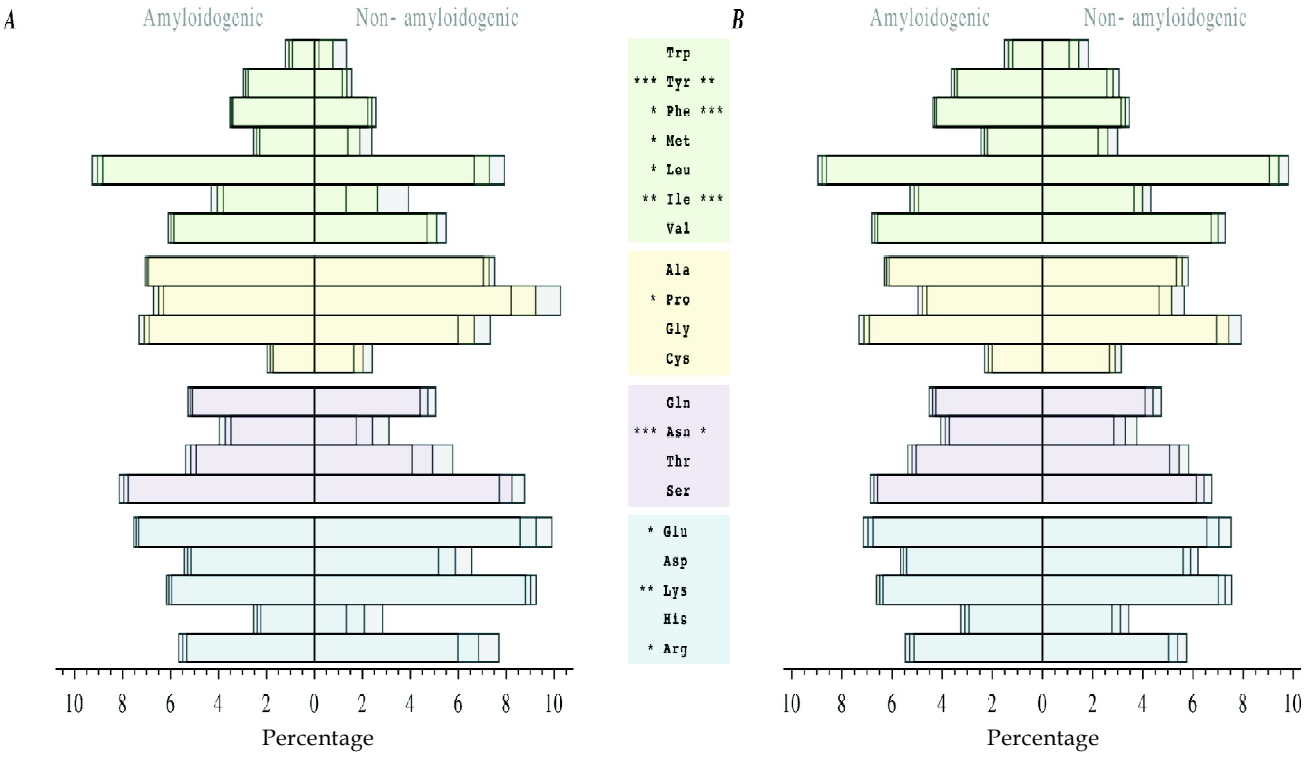

Figure 5: **Comparison of amino acid composition between amyloidogenic and non-amyloidogenic disordered (A) and structured (B) human proteins.**
Amino acids were categorized in four groups and shown bottom-up in the following order: charged (light blue), uncharged polar (light purple), special types (light yellow) and hydrophobic (light green). Significant variations are marked with asterisks. The symmetry was most effected in the disordered group.

hydrophobicity in disordered amyloidogenic proteins (Figure 2). The differences in the charged amino acid composition also contributed in part to the pI distribution of the amyloidogenc and non-amyloidogenic disordered proteins (Figure 1).

Conclusion

Understanding the factors, governing the protein stability in solution state is a key step to apprehend the protein solubility, aggregation propensity and misfolding related diseases such as Alzheimer's, Parkinson's diseases and Prion disorder. In this work, we discussed several physicochemical parameters that govern the stability and aggregation behavior of amyloidogenic and non-amyloidogenic human proteins. The exact mechanism of amyloid formation and how protein aggregation exerts toxic effect towards different cellular systems are very obscure. One hypothesis is that a protein needs to contain intrinsically disordered region or in case of a structured protein, it needs to partially or fully unfold to form amyloid fiber. In this regard, amino acid composition and related physicochemical properties are very important parameters for investigation. The results indicated that the sequence composition and the physicochemical properties such as pI, GRAVY, AI and II of amyloidogenic and non-amyloidogenic proteins differed in large extent, but almost exclusively for the intrinsically disordered proteins. It also explains from this perspective why a drastic solution condition is required for a structured protein to form amyloid aggregates even though it contains many ARs. Our findings could be useful in the development of better amyloidogenic protein detection algorithms.

Acknowledgements

Uttam Pal and Swagata Das thanks INSPIRE Fellowship Programme, Department of Science and Technology, Government of India, India for financial support. Supriya Das and Anupam Roy thanks University Grants Commission, Government of India, India for financial support. We also acknowledge the grant supports (project heads BSC-0115, BSC-0113, BSC-0121 and GAP-299) from the CSIR and the Department of Biotechnology, government of India.

Supporting information

Supplementary data: It contains the Tables S1 and S2 enlisting information about all the proteins in the dataset that were used for the analysis. It also contains the Figure S1 and Table S3.

Conflict of Interest

The authors do not have any conflict of interest with the contents of this manuscript.

References

Anderson, T.W. and Darling, D.A. (1952). Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. Ann. Math. Stat. *23*, 193–212.

Belli, M., Ramazzotti, M. and Chiti, F. (2011). Prediction of amyloid aggregation in vivo. EMBO Rep. *12*, 657–663.

Bemporad, F., Calloni, G., Campioni, S., Plakoutsi, G., Taddei, N. and Chiti, F. (2006). Sequence and Structural Determinants of Amyloid Fibril Formation. Acc. Chem. Res. *39*, 620–627.

Calamai, M., Taddei, N., Stefani, M., Ramponi, G. and Chiti, F. (2003). Relative Influence of Hydrophobicity and Net Charge in the Aggregation of Two Homologous Proteins†. Biochemistry (Mosc.) *42*, 15078–15083.

Chaitanya, N.K., Paul, A., Saha, A. and Mandal, B. (2013). Modulation of Aggregation Propensity of Aâ38 by Site Specific Multiple Proline Substitution. Int. J. Pept. Res. Ther. *19*, 365–371.

Chennamsetty, N., Voynov, V., Kayser, V., Helk, B. and Trout, B.L. (2010). Prediction of Aggregation Prone Regions of Therapeutic Proteins. J. Phys. Chem. B *114*, 6614–6624.

Chiti, F. and Dobson, C.M. (2009). Amyloid formation by globular proteins under native conditions. Nat. Chem. Biol. *5*, 15–22.

Chiti, F., Taddei, N., Baroni, F., Capanni, C., Stefani, M., Ramponi, G., and Dobson, C.M. (2002). Kinetic partitioning of protein folding and aggregation. Nat. Struct. Mol. Biol. *9*, 137–143.

Chiti, F., Stefani, M., Taddei, N., Ramponi, G. and Dobson, C.M. (2003). Rationalization of the effects of mutations on peptide and protein aggregation rates. Nature *424*, 805–808.

Das, S., Pal, U., Das, S. and Maiti, N.C. (2013). Chaperone action of cyclophilin on lysozyme and its aggregate. J. Proteins Proteomics *4*, 129.

Das, S., Pal, U., Das, S., Bagga, K., Roy, A., Mrigwani, A. and Maiti, N.C. (2014). Sequence complexity of amyloidogenic regions in intrinsically disordered human proteins. PLoS ONE *9*, e89781.

Dosztányi, Z., Chen, J., Dunker, A.K., Simon, I. and Tompa, P. (2006). Disorder and Sequence Repeats in Hub Proteins and Their Implications for Network Evolution. J. Proteome Res. *5*, 2985–2995.

Fändrich, M., Fletcher, M.A. and Dobson, C.M. (2001). Amyloid fibrils from muscle myoglobin. Nature *410*, 165–166.

Fang, Y. and Fang, J. (2013). Discrimination of soluble and aggregation-prone proteins based on sequence information. Mol. Biosyst. *9*, 806.

Ferron, F., Longhi, S., Canard, B. and Karlin, D. (2006). A practical overview of protein disorder prediction methods. Proteins Struct. Funct. Bioinforma. *65*, 1–14.

Galea, C.A., High, A.A., Obenauer, J.C., Mishra, A., Park, C.-G., Punta, M., Schlessinger, A., Ma, J., Rost, B., Slaughter, C.A., *et al.* (2009). Large-Scale Analysis of Thermostable, Mammalian Proteins Provides Insights into the Intrinsically Disordered Proteome. J. Proteome Res. *8*, 211–226.

Garbuzynskiy, S.O., Lobanov, M.Y. and Galzitskaya, O.V. (2010). FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. Bioinformatics *26*, 326–332.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S. 'everine, Wilkins, M.R., Appel, R.D. and Bairoch, A. (2005). Protein Identification and Analysis Tools on the ExPASy Server. In The Proteomics Protocols Handbook, J.M. Walker, ed. (Humana Press), pp. 571–607.

de Groot, N.S., Aviles, F.X., Vendrell, J. and Ventura, S. (2006). Mutagenesis of the central hydrophobic cluster in Aβ42 Alzheimer's peptide. FEBS J. *273*, 658–668.

Guruprasad, K., Reddy, B.V.B. and Pandit, M.W. (1990). Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. Protein Eng. *4*, 155–161.

Idicula-Thomas, S. and Balaji, P.V. (2005). Understanding the relationship between the primary structure of proteins and their amyloidogenic propensity: clues from inclusion body formation. Protein Eng. Des. Sel. *18*, 175–180.

Ikai, A. (1980). Thermostability and Aliphatic Index of Globular Proteins. J. Biochem. (Tokyo) *88*, 1895–1898.

Kiraga, J., Mackiewicz, P., Mackiewicz, D., Kowalczuk, M., Biecek, P., Polak, N., Smolarczyk, K., Dudek, M.R. and Cebrat, S. (2007). The relationships between the isoelectric point and: length of proteins, taxonomy and ecology of organisms. BMC Genomics *8*, 163.

Kovačević, J.J. (2012). Computational analysis of position-dependent disorder content in DisProt database. Genomics Proteomics Bioinformatics *10*, 158–165.

Maity, M. and Maiti, N.C. (2012). Sequence Composition of Binding Sites in Natively Unfolded Human Proteins. J. Proteins Proteomics *3*, 117–125.

Maurer-Stroh, S., Debulpaep, M., Kuemmerer, N., Paz, M.L. de la, Martins, I.C., Reumers, J., Morris, K.L., Copland, A., Serpell, L., Serrano, L., *et al.* (2010). Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. Nat. Methods *7*, 237–242.

McParland, V.J., Kad, N.M., Kalverda, A.P., Brown, A., Kirwin-Jones, P., Hunter, M.G., Sunde, M. and Radford, S.E. (2000). Partially Unfolded States of â2-Microglobulin and Amyloid Formation in Vitro†. Biochemistry (Mosc.) *39*, 8735–8746.

Meuvis, J., Gerard, M., Desender, L., Baekelandt, V, and Engelborghs, Y. (2010). The Conformation and the Aggregation Kinetics of á-Synuclein Depend on the Proline Residues in Its C-Terminal Region. Biochemistry (Mosc.) *49*, 9345–9352.

Monsellier, E., Ramazzotti, M., Taddei, N. and Chiti, F. (2008). Aggregation Propensity of the Human Proteome. PLoS Comput Biol *4*, e1000199.

Morimoto, A., Irie, K., Murakami, K., Ohigashi, H., Shindo, M., Nagao, M., Shimizu, T. and Shirasawa, T. (2002). Aggregation and neurotoxicity of mutant amyloid â (Aβ) peptides with proline replacement: importance of turn formation at positions 22 and 23. Biochem. Biophys. Res. Commun. *295*, 306–311.

Nielsen, L., Khurana, R., Coats, A., Frokjaer, S., Brange, J., Vyas, S., Uversky, V.N. and Fink, A.L. (2001). Effect of Environmental Factors on the Kinetics of Insulin Fibril Formation: Elucidation of the Molecular Mechanism†. Biochemistry (Mosc.) *40*, 6036–6046.

Oldfield, C.J., Cheng, Y., Cortese, M.S., Brown, C.J., Uversky, V.N. and Dunker, A.K. (2005). Comparing and Combining Predictors of Mostly Disordered Proteins†. Biochemistry (Mosc.) *44*, 1989–2000.

Pal, U., Maity, M., Khot, N., Das, S., Das, S., Dolui, S. and Maiti, N.C. (2016). Statistical insight into the binding regions in disordered human proteome. J. Proteins Proteomics *7*, 47.

Pawar, A.P., DuBay, K.F., Zurdo, J., Chiti, F., Vendruscolo, M. and Dobson, C.M. (2005). Prediction of "Aggregation-prone" and "Aggregation-susceptible" Regions in Proteins Associated with Neurodegenerative Diseases. J. Mol. Biol. *350*, 379–392.

Routledge, K.E., Tartaglia, G.G., Platt, G.W., Vendruscolo, M. and Radford, S.E. (2009). Competition between Intramolecular and Intermolecular Interactions in an Amyloid-Forming Protein. J. Mol. Biol. *389*, 776–786.

Sickmeier, M., Hamilton, J.A., LeGall, T., Vacic, V., Cortese, M.S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V.N., *et al.* (2007). DisProt: the Database of Disordered Proteins. Nucleic Acids Res. *35*, D786–D793.

Taylor, J.P., Hardy, J. and Fischbeck, K.H. (2002). Toxic Proteins in Neurodegenerative Disease. Science *296*, 1991–1995.

Tcherkasskaya, O., Davidson, E.A. and Uversky, V.N. (2003). Biophysical Constraints for Protein Structure Prediction. J. Proteome Res. *2*, 37–42.

Vernaglia, B.A., Huang, J. and Clark, E.D. (2004). Guanidine Hydrochloride Can Induce Amyloid Fibril Formation from Hen Egg-White Lysozyme. Biomacromolecules *5*, 1362–1370.

Westermark, P., Benson, M.D., Buxbaum, J.N., Cohen, A.S., Frangione, B., Ikeda, S.-I., Masters, C.L., Merlini, G., Saraiva, M.J. and Sipe, J.D. (2005). Amyloid: Toward

terminology clarification Report from the Nomenclature Committee of the International Society of Amyloidosis. Amyloid *12*, 1–4.

Wootton, J.C. (1994). Sequences with "unusual" amino acid compositions. Curr. Opin. Struct. Biol. *4*, 413–421.

Wootton, J.C. and Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. Comput. Chem. *17*, 149–163.

Wu, S., Wan, P., Li, J., Li, D., Zhu, Y. and He, F. (2006). Multi-modality of pI distribution in whole proteome. PROTEOMICS *6*, 449–455.

Xie, H., Vucetic, S., Iakoucheva, L.M., Oldfield, C.J., Dunker, A.K., Uversky, V.N. and Obradovic, Z. (2007a). Functional Anthology of Intrinsic Disorder. 1. Biological Processes and Functions of Proteins with Long Disordered Regions. J. Proteome Res. *6*, 1882–1898.

Xie, H., Vucetic, S., Iakoucheva, L.M., Oldfield, C.J., Dunker, A.K., Obradovic, Z. and Uversky, V.N. (2007b). Functional Anthology of Intrinsic Disorder. 3. Ligands, Post-Translational Modifications, and Diseases Associated with Intrinsically Disordered Proteins. J. Proteome Res. *6*, 1917–1932.

Zbilut, J.P., Giuliani, A., Colosimo, A., Mitchell, J.C., Colafranceschi, M., Marwan, N., Webber, Charles L. and Uversky, V.N. (2004). Charge and Hydrophobicity Patterning along the Sequence Predicts the Folding Mechanism and Aggregation of Proteins: A Computational Approach. J. Proteome Res. *3*, 1243–1253.