doi:10.1016/j.jmb.2004.06.088 *J. Mol. Biol.* (2004) **342**, 345–353

JMB

Available online at www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

A Comparative Study of the Relationship Between Protein Structure and β-Aggregation in Globular and Intrinsically Disordered Proteins

Rune Linding[1]†, Joost Schymkowitz[2]†, Frederic Rousseau[2]† Francesca Diella[1] and Luis Serrano[1]*

[1]*European Molecular Biology Laboratory, Programme for Structural and Computational biology, Meyerhofstr 1, D-69117 Heidelberg, Germany*

[2]*Flanders Interuniversity Institute of Biotechnology (VIB) SWITCH laboratory, Vrije Universiteit Brussel, Pleinlaan 2 1050, Brussels, Belgium*

**Corresponding author*

A growing number of proteins are being identified that are biologically active though intrinsically disordered, in sharp contrast with the classic notion that proteins require a well-defined globular structure in order to be functional. At the same time recent work showed that aggregation and amyloidosis are initiated in amino acid sequences that have specific physico-chemical properties in terms of secondary structure propensities, hydrophobicity and charge. In intrinsically disordered proteins (IDPs) such sequences would be almost exclusively solvent-exposed and therefore cause serious solubility problems. Further, some IDPs such as the human prion protein, synuclein and Tau protein are related to major protein conformational diseases. However, this scenario contrasts with the large number of unstructured proteins identified, especially in higher eukaryotes, and the fact that the solubility of these proteins is often particularly good. We have used the algorithm TANGO to compare the β aggregation tendency of a set of globular proteins derived from SCOP and a set of 296 experimentally verified, non-redundant IDPs but also with a set of IDPs predicted by the algorithms DisEMBL and GlobPlot. Our analysis shows that the β-aggregation propensity of all-α, all-β and mixed α/β globular proteins as well as membrane-associated proteins is fairly similar. This illustrates firstly that globular structures possess an appreciable amount of structural frustration and secondly that β-aggregation is not determined by hydrophobicity and β-sheet propensity alone. We also show that globular proteins contain almost three times as much aggregation nucleating regions as IDPs and that the formation of highly structured globular proteins comes at the cost of a higher β-aggregation propensity because both structure and aggregation obey very similar physico-chemical constraints. Finally, we discuss the fact that although IDPs have a much lower aggregation propensity than globular proteins, this does not necessarily mean that they have a lower potential for amyloidosis.

Keywords: intrinsic protein disorder; hot loops; structural proteomics; aggregation; amyloidosis

Introduction

Protein aggregation has long been thought of as an unspecific process caused by the formation of non-native contacts between protein folding intermediates. Recent work, however, shows that often aggregation is a much more specific process than previously expected and that, accordingly, it can be reliably correlated to a combination of simple physico-chemical parameters.[1–3] In particular, several models for aggregation were postulated that all involve the formation of an intermolecular β-sheet initiated by amino acid sequences that act as nuclei for β-aggregation.[4–8] According to these models, aggregation is initiated when amino acid segments having a high hydrophobicity, a good

† R.L., J.S. and F.R. contributed equally to this work.

Abbreviations used: IDPs, intrinsically disordered proteins; PDB, Protein Data Bank.

E-mail address of the corresponding author: serrano@embl-heidelberg.de β-sheet propensity and a low net charge are solvent-exposed so that they can associate. As a result one would then expect aggregating protein segments to be buried in the folded state and not to be exposed to the solvent. This is confirmed by the experimental finding that in many globular proteins, aggregation occurs during refolding or under conditions in which denatured or partially folded states are significantly populated, i.e. at high concentration or as a result of destabilizing conditions or mutations.[9] Based on these findings we recently developed the computer programme TANGO[10] to predict β-aggregating stretches in proteins, based on a statistical mechanics algorithm that considers the physico-chemical parameters described above but also competition between different structural conformations: β-turn, α-helix, β-sheet aggregates and the folded state. The algorithm is based on the assumption that in the ordered β-aggregates the nucleating regions end up fully buried, paying maximal desolvation energy as well as entropy, while satisfying their H-bonding potential. The energy contributions are derived from the FOLD-X force field.[11] In a blind test involving 174 peptides from over 20 proteins, TANGO achieved an accuracy of 95% in predicting aggregating sequences, as well as the effect of point mutations on the aggregation tendency of proteins.[10] Many intrinsically disordered proteins (IDPs) have been discovered in all kingdoms of life, but especially in higher eukaryotes.[12–14] These are proteins or domains that, in their native state, are either completely disordered or contain large disordered regions.[15,16] More than 180 such proteins are known to date, including prions, CREB, Tau, MAPs and p53.[16] These polypeptides perform important regulatory functions and are widespread in eukaryotic cells and tissue. Some acquire structure upon binding to another protein or DNA, others act as structural anchors in large protein–protein and protein–RNA complexes, making use of extended interaction surfaces that are simply not available in more compact conformations.[12] Furthermore, many globular proteins contain disordered segments acting as functional modules, e.g. post-translational modification sites and domain ligands. Importantly, many IDPs are involved in key cellular processes and some of them are related to major protein conformational diseases, e.g. prions (BSE), Tau (Alzheimer's disease), and synuclein (Parkinson's disease). The uniting factor associating the above proteins to their disease states is a high degree of aggregation or amylogenicity. Amylogenicity is not itself a direct result of β-aggregation but it is often found in association with and can be strongly promoted by β-aggregation.[17] On the other hand, as mentioned above, it is often found experimentally that unstructured proteins are resistant to aggregation, even under harsh treatments such as incubation at high temperature.[16] In fact, heat-exposure of cell-extracts is an effective protocol for purification of several recombinantly expressed unstructured proteins.[16] It is therefore important to investigate the relationship between intrinsic disorder and aggregation to gain further insight into the potential of IDPs to be implicated into protein conformational diseases. The TANGO algorithm offers the opportunity to compare the aggregation propensities of IDPs and globular proteins, not only by considering average aggregation-related physico-chemical properties, but also by directly comparing the nature and frequency of aggregation-promoting nucleation stretches. This analysis should therefore allow us to test whether disorder does correlate with aggregation, as some cases of disease association suggest, or whether it anti-correlates with aggregation as residue compositional biases of IDPs suggest.

In order to deal with this issue we have used TANGO to compare the aggregation tendency of a non-redundant set of globular proteins derived from the SCOP database (the ASTRAL40 set, see Materials and Methods),[18] a set of proteins that were experimentally shown to be unstructured[16,19] as well as a set of predicted disordered protein sequences. Data sets of experimentally verified disordered proteins are scarce and rather error-prone, hence we have collected and cured a set of 296 experimentally verified and published, IDP sequences. This is to our knowledge the largest dataset available to the community. The datasets of predicted disordered segments or proteins were predicted by the DisEMBL[20] and GlobPlot[21] algorithms and divided into sequences of low (∼50%) and average sequence complexity.

Our analysis clearly shows that aggregation-prone segments are much less frequent in IDPs than in globular proteins, thus accounting for their good solubility. Although more frequent in globular proteins, β-aggregating segments are generally part of the hydrophobic core. These observations show that the compositional bias observed in IDPs reduces secondary and tertiary structure as well as aggregation because both structure and aggregation rely on similar physico-chemical properties. As previously observed,[12,16] IDPs are not completely devoid of structure, as should be expected if some degree of functional specificity has to be obtained, but they perform their particular cellular functions by achieving a low degree of order, retaining only structural propensities that are devoid of aggregation-promoting features.

Results and Discussion

TANGO score for aggregation and accuracy of the TANGO algorithm

The TANGO algorithm was calibrated using data found in the scientific literature on the aggregation of 174 peptides corresponding to sequence fragments of 21 different proteins, studied by various research groups using circular dichroism (CD) or nuclear magnetic resonance (NMR). Of the peptides

Table 1. Comparison of the aggregation tendency in various datasets

	Aggregation-prone regions per 1000 residues	Number of protein sequences in dataset
Globular proteins ASTRAL 40		
All-α proteins	18	129
All-β proteins	20	174
α and β proteins (α/β)	18	190
α and β proteins (α+β)	19	162
Membrane and cell-surface proteins	16	11
Multi-domain proteins	13	25
Experimentally validated intrinsically disordered proteins		
IDPs, experimentally validated	7	183
IDPs, experimentally validated, low complexity	4	27
Predicted intrinsically disordered proteins		
IDPs defined by Russell/Linding	6	602
IDPs defined by Russell/Linding, low complexity	5	60
IDPs predicted by Hot-loops	8	7925
IDPs predicted by Hot-loops, low complexity	3	59
IDPs predicted by Coils/Loops	8	16,988
IDPs predicted by Coils/Loops, low complexity	4	74
IDPs predicted by Remark-465	6	623
IDPs predicted by Remark-465, low complexity	4	139

See Materials and Methods for details of the dataset composition.

in our set, 70 were experimentally observed to aggregate in the concentration range between 100 μM and 1 mM, while the others remained soluble in this concentration range. A detailed description of our results together with the source of the experimental data is given by Fernandez-Escamilla *et al.*[10]. For each residue in a peptide TANGO returns its percentage occupancy of the β-aggregation conformation. We consider a peptide to possess a β-aggregation nucleation site when it has a segment of at least five consecutive residues, each populating the β-aggregated conformation for more than 5%. Using this threshold level the overall accuracy of TANGO reaches 95% compared with our experimental set.

Statistical considerations

In order to exclude a bias in our results due to the difference in number of proteins in the globular and IDP sets or due to the average sequence length of IDP and globular proteins, 25 control sets were generated by randomly selecting 10,000 sequence fragments of 20 amino acid residues from each original set. These were analysed in exactly the same way as the original datasets and the standard deviation between the results of original and control sets was calculated. The standard deviation was below one window per 1000 residues, indicating that our results are not strongly influenced by the composition and size of the sets. Further, to exclude bias due to the over-representation of certain sequences, our sequence sets for IDPs and globular proteins were reduced in identical fashion so that a maximum of 40% sequence similarity remained between any pair of sequences within the set.

The frequency of β-aggregation nucleation sites in different structural classes of globular proteins is very similar

In Table 1 we show the aggregation tendency for globular proteins of different fold classes. Interestingly, we do not find a significant difference in the occurrence of β-aggregating nucleation regions between the different fold classes (all-α, all-β, α/β and α+β), including membrane proteins. Thus in globular proteins nucleating regions for β-aggregation are as likely to occur in α-helices as in β-sheets, demonstrating a certain degree of structural frustration in protein structures.[22,23] The fact that we cannot distinguish membrane proteins that contain large hydrophobic trans-membrane segments from cytoplasmic proteins was to be expected as β-aggregation tendency does not only rely on hydrophobicity but is equally determined by entropic, H-bonding and electrostatic factors.[10] In order to be a β-aggregating segment a residue segment must accumulate several physico-chemical characteristics, such as a high hydrophobicity, a good β-sheet propensity, complementarity of H-bonding potential and charges and a low overall net charge. The sequence MAMAMAMA, for instance, although being quite hydrophobic, will have only a moderate β-aggregation tendency (12%) because it is entropically unfavourable, having a low β-sheet propensity and a high α-helical tendency, while VAVAVAVA, on the contrary, will have a very strong aggregation propensity (99%), since it combines a high hydrophobicity with a strong β-sheet propensity. Replacing the Ala by Thr increases the aggregation tendency in the first sequence and only slightly decreases aggregation in the second sequence (from 12% to 31% and from 99% to 83%, respectively), even though Thr is a polar residue, because it can satisfy its H-bonding potential with itself. Even introducing two oppositely charged residues in the VAVAVAVA sequence at positions 4 and 6, does not decrease the aggregation too much (76%), because of the charge complementarity.

Thus although hydrophobicity is an important factor, it can be modulated by many other contributions[10] and therefore the aggregation tendency cannot be equated with hydrophobic sequences.

β-Aggregation nucleation segments occur with a much higher frequency in globular proteins than in disordered proteins

Figure 1 shows the number of β-aggregation nucleating segments as defined above per 1000 amino acid residues. The frequency of β-aggregation nucleating segments is much higher in globular proteins than in IDPs or in unstructured regions of globular proteins. TANGO detects about 18 aggregation nucleating regions per 1000 residues in globular proteins, whereas this number drops to seven in IDPs. Furthermore, 70% of IDPs do not possess an aggregation nucleation segment, while only 20% of the ASTRAL40 database entries fulfil the same criterion (Figure 2). Incidentally, the TANGO score for the experimentally verified IDPs and for the set of IDPs predicted by DisEMBL/GlobPlot are very similar, underlining the accuracy of these algorithms. A combined webserver† provides an integrated interface for TANGO and DisEMBL[20] algorithms, predicting aggregation nucleation sites in unstructured proteins or unstructured regions of proteins. β-Aggregation nucleating segments are typically enriched with β-branched hydrophobic residues such as valine and isoleucine but also aromatic residues such as tryptophan and phenylalanine, and are depleted of net charge and proline residues. The effect of the bias in the amino acid composition observed in IDP sequences is therefore not only to reduce structure by lowering tertiary structure as well as β-sheet propensity and to a lesser extent α-helical propensities, but also to reduce β-aggregation propensities, because both globular structure and β-aggregation have overlapping physico-chemical constraints.

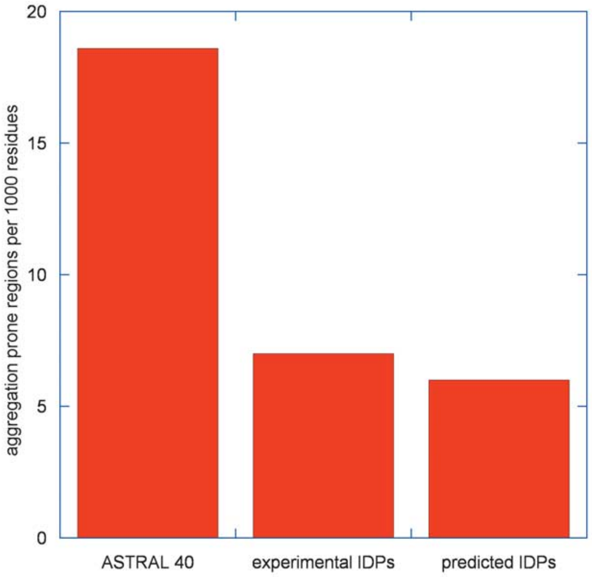

Figure 1. Comparison of aggregation tendency in globular (ASTRAL40) and unstructured proteins (IDPs): overview. For details, see Table 1.

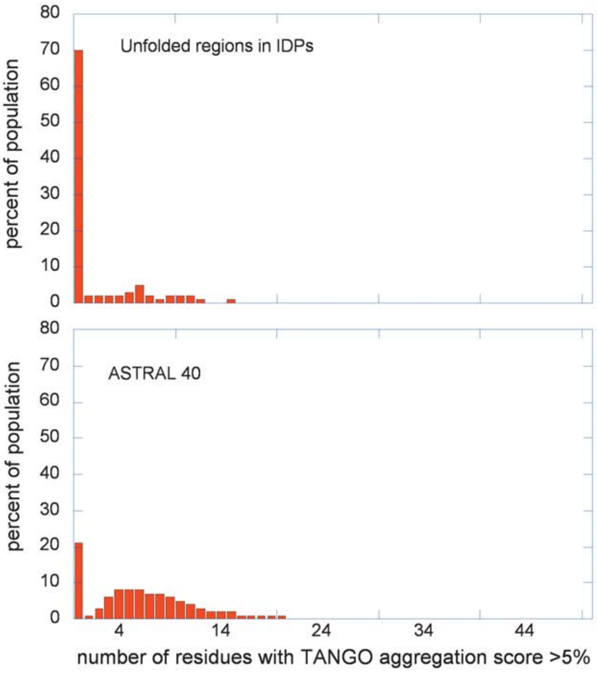

Figure 2. Distribution of aggregation tendency in datasets: number of residues that have a TANGO score larger than 5%, per 1000 residues. Upper panel, 70% of unfolded regions in experimentally validated IPDs do not contain aggregation-prone regions. Lower panel, only 20% of ASTRAL40 proteins do not contain aggregation-prone stretches.

β-Aggregation nucleating regions in globular proteins are buried in the hydrophobic core

As discussed above, globular domains have a significantly higher number of aggregating regions, on average, than IDPs. Further, only 20% of all globular proteins in the ASTRAL40 set have no aggregation nucleating regions, while 70% of the IDPs are devoid of it. More than 50% of the polypeptide backbone of globular proteins adopts an α-helical or β-sheet conformation that packs together forming the often extensive hydrophobic core that maintains a well-defined tertiary structure. To achieve this the amino acid composition of globular proteins must be rather rich in hydrophobic residues, preferably β-branched hydrophobic residues that can easily fit in extended β-strand conformations but also in α-helices. This

† http://dis.embl.de; http://globplot.embl.de

will automatically imply a higher degree of β-aggregation, since β-aggregation is also favoured by hydrophobic, β-branched residues. Therefore, the question in fact is how proteins achieve globular structures without aggregating massively. In Figure 3 we show the solvent accessibility of all the aggregation regions found in the ASTRAL40 database. It is quite clear that, with few exceptions, aggregation-prone regions have a low solvent accessibility. Therefore in globular proteins aggregation nucleating regions are generally buried in the hydrophobic core and as such kept protected from solvent by a dense network of cooperative interactions stabilizing the native state. Only when the protein is unfolded, as it is when it is synthesized on the ribozome or destablized by mutation, change of temperature or pH, thereby favouring the unfolded state, will it aggregate. Even so, globular proteins have less aggregating segments than hydrophobic secondary structure elements, indicating that there is a selection against aggregating sequences.[17]

β-Aggregation and disease in globular proteins and IDPs

What is the role of β-aggregation in protein conformational diseases, what is its relation with amyloidosis and cytotoxicity, and does it play an equal role both in globular proteins and IDPs? In recent years it became clear that amyloidosis is a conformational process not only associated with specific disease-related proteins, but that it can also be induced in non-disease-related proteins.[1,24] Further, in most instances the precursor aggregates during amyloid formation are often more toxic than the amyloid fibres themselves. The mechanism of toxicity of these precursors seems to be universal and to involve precursor–cell membrane interactions that trigger changes in the concentration of intracellular free calcium followed by apoptosis.[25,26] Finally, the structure of these toxic precursor aggregates is invariably rich in β-strands and must be very similar indeed, as antibodies specifically raised against Aβ-peptide amyloid precursors can bind amyloid precursors of other proteins with completely different sequences.[27] Here, using TANGO, we analysed the relationship of β-aggregation with amyloidosis and cytotoxicity in some amylogenic IDPs and globular proteins. As seen from Figure 4, two globular disease-related amylogenic proteins (human lysozyme and β-microglobulin) and two disease-related IDPs (Alzheimer's β-peptide and α-synuclein) also possess β-aggregation nucleation segments according to TANGO. Further, in Figure 4 we also analysed the relationship between β-aggregation and amyloidosis in two yeast prion proteins (sup35p, ure2p). These highly amylogenic IDPs are not toxic for yeast and it has been suggested that in those and other similar proteins amyloidosis could play a functional role.[28] Interestingly, these proteins, although being amyloidogenic, do not possess β-aggregation nuclei. These observations show that although β-aggregation can promote amyloidosis, β-aggregation and amyloidosis are not necessarily coupled, as some proteins form β-aggre-aggregates without forming amyloid fibres, while others are amylogenic without possessing strong β-aggregation nucleation segments.[17] The co-occurrence of β-aggregation and cytotoxicity in the disease-related examples above, as well as the fact that amyloid precursor aggregates are often more toxic than the amyloids themselves, could suggest that in fact disease correlates with β-aggregation propensity and not amyloidosis *per se*. However, when comparing the β-aggregation tendency, cytotoxicity and amylogenicity of PI3-SH3 and α-spectrin SH3 we note that PI3-SH3 does β-aggregate, forms amyloid fibres and is cytotoxic.[29,30] α-Spectrin SH3, on the other hand has a higher aggregating tendency than PI3-SH3[31] but does not form fibres under any conditions and neither is it cytotoxic.[32] As a conclusion, these examples show that although the conditions causing β-aggregation, amyloidosis and cytotoxicity partly overlap, they are not strictly dependent on each other. The observation of a compositional bias disfavouring aggregation therefore does not necessarily imply a lower propensity for amyloidosis or a lower potential cytotoxicity in IDPs.

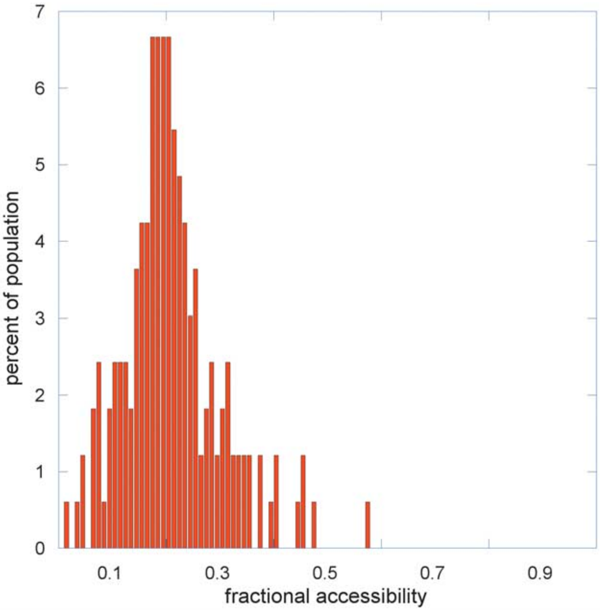

Figure 3. Histogram showing the average solvent accessibility of aggregation-prone regions to be low. The solvent-accessible surface area of the aggregation-prone regions (defined as segments of five consecutive residues with a TANGO score of more than 5% per residue) in the ASTRAL40[18] dataset was calculated with the program DSSP.[39]

Conclusions

TANGO is an algorithm to predict β-aggregation

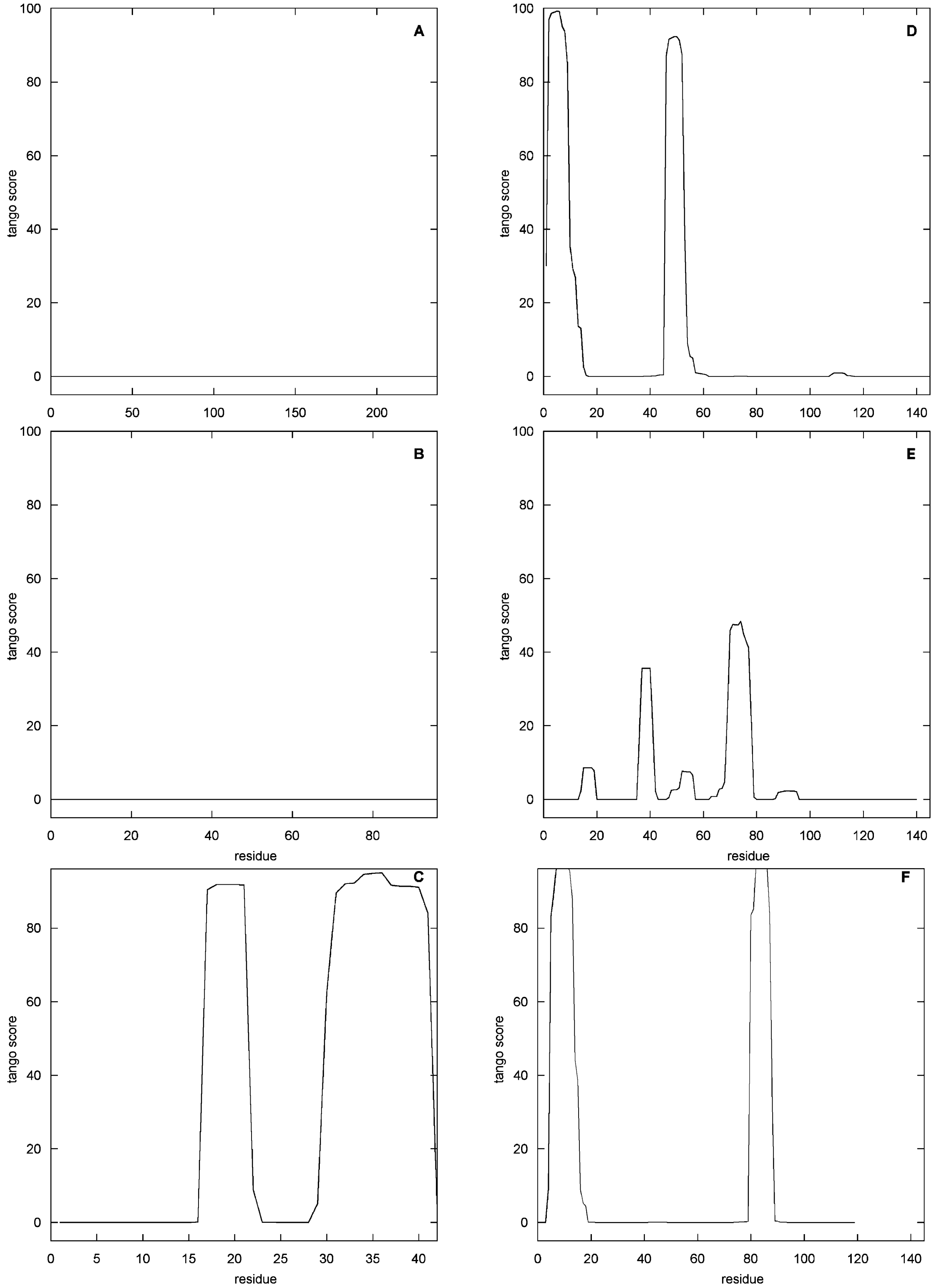

Figure 4. TANGO output per residue for the following proteins. A, sup35; B, ure2; C, Alzheimer β-peptide; D, human lysozyme; E, α-synuclein; and F, β-microglobulin.

nucleating regions in proteins. Here, we used TANGO to compare the β-aggregation propensities of globular proteins and intrinsically unstructured proteins. In globular proteins we found similar amounts of β-aggregating nucleation regions in all-α, all-β and mixed α/β proteins. This demonstrates that globular proteins do display a certain degree of structural frustration and can at the same time display propensities for both α and β conformations under destabilizing conditions even when the native state is α-helical. Further, membrane-associated proteins do not possess a higher frequency of β-aggregation nucleating segments within their sequences than cytoplasmic proteins. The reason for this lies in the fact that β-aggregation is not only determined by hydrophobicity but also by entropic factors, hydrogen bonding and electrostatics. Next we observe that the frequency of aggregation nucleating segments is almost three times lower in IDPs than in globular proteins, dropping from 18 segments per 1000 residues on average for globular proteins to seven segments per 1000 residues for IDPs. Further, 70% of the IDPs are devoid of aggregation nucleating segments, while only 20% of globular proteins are devoid of it. However, the vast majority of aggregation nuclei in globular proteins are buried inside the hydrophobic core and protected from aggregation by the cooperative interactions stabilizing the native state. The higher intrinsic aggregation propensity of globular proteins is directly related to the physico-chemical requirements to form globular structures. Because those requirements overlap with the conditions favouring aggregation, formation of globular structures goes at the cost of a higher aggregation propensity. Finally, we show that although β-aggregation propensity, amyloidosis and cytotoxicity are often related to each other, this does not necessarily mean that the lower aggregation propensity in IDPs implies a lower potential for amyloidosis (as illustrated by yeast prion proteins) or cytotoxicity (β-Alzheimer peptide).

Materials and Methods

Datasets

Here we have used datasets that cover globular and IDPs. Both predicted and experimentally verified datasets are described and for each dataset we have split the data into a low and a normal-complexity set.

Experimentally verified IDPs

Databases of intrinsically disordered proteins are error-prone and rather scarce. We have built a cured and expanded database of 296 experimentally verified proteins. This dataset was annotated from the literature[16,19] and from the Dunker laboratory data†. The 40% non-redundant subset consists of 183 non-homologous sequences. Redundancy reduction was done by the CD-HI algorithm.[33] All disordered segments were mapped into the SWISS-PROT[34] database to get full-length chains. This ensures a more precise disorder/order residue count within the set. This dataset is available‡.

Predicted IDPs

No commonly agreed definition of protein disorder exists. DisEMBL[20] and GlobPlot[21] provide four different definitions of disorder.

The DisEMBL disorder definitions are known as Coils/Loops, Remark-465 (missing X-ray data) and Hot-loops. GlobPlots primary disorder definition is known as Russell/Linding disorder.[20,21] We have predicted IDPs according to each of these four definitions. In all cases a minimum disorder length of 50 consecutive residues was used to identify putative IDPs. All the predictions were done in a 40% similarity reduced human proteome set. The proteome was based on the EBI§ proteome cross-checked with the mouse predictions to reduce erroneous gene predictions.

Globular proteins

To represent globular proteins we used the Astral 1.63 40% subset of SCOP[18,35] (where 40% indicates the upper limit of sequence identity between members in the dataset).

Low-complexity

All sequences were subsequently classified according to low-complexity as determined by CAST.[36] Only the segments predicted to be disordered by DisEMBL or GlobPlot were analysed with CAST. If at least one disordered segment contained >50% low-complexity residues the sequence was categorised as a low-complexity IDP.

The TANGO model

The model used by the TANGO algorithm is designed to predict β-aggregation in peptides and denatured proteins and consists of a phase-space encompassing the random coil and possible structural states: β-turn, α-helix, β-sheet aggregation and the folded state in the case of globular proteins. Every segment of a peptide can populate each of these states (except the folded state that encompasses the whole sequence) according to a Boltzmann distribution, i.e. the frequency of population of each structural state for a given segment will be relative to its energy. Therefore, to predict β-aggregating segments of a peptide TANGO simply calculates the partition function of the phase-space. Here, we give an overview of the main features of TANGO, a detailed description of the algorithm will be given elsewhere (A. M. Escamilla-Fernandez, F.R., J.S. & L.S., unpublished results).

α-Helical propensities

The parameters used in the latest version of AGADIR (AGADIR-1s[37,38]), have been used to determine the helical propensity of the amino acid sequences. The only modification has been the implementation of a two-window approximation (A. M. Escamilla-Fernandez, F.R., J.S. & L.S., unpublished results).

β-Turn propensities

β-Turn propensity is calculated by considering three energy contributions: (1) an amino acid-specific cost in

† http://divac.ist.temple.edu/disprot

‡ http://dis.embl.de/datasets/idp40.fas.gz

§ http://www.ebi.ac.uk/proteome/

conformational entropy for fixing that residue in a β-turn compatible conformation; (2) interactions of each amino acid with the turn structure in a position-dependent manner; and (3) a single H-bond between the main chains of residues i and $i+3$ of the turn. We have only considered four types of turns for which we could obtain significant statistical data, types I, I′, II and II′.

β-Sheet aggregation

To estimate the aggregation tendency of a particular amino acid sequence, we have taken the following assumptions: (1) in an ordered β-sheet aggregate the main secondary structure is the β-strand. (2) The regions involved in the aggregation process are fully buried, thus paying full solvation costs and gains, full entropy and optimized H-bond potential (i.e. the number of H-bonds made in the aggregate is related to the number of donor groups that are compensated by acceptors; an excess of donors or acceptors remains unsatisfied). (3). Complementary charges in the selected window establish favourable electrostatic interactions and overall net charge of the peptide inside and outside the window disfavours aggregation.

The effect of physico-chemical conditions on aggregation

The effect of pH, temperature and ionic strength on electrostatic interactions was taken into account as described in AGADIR2-1s.[37,38] Similarly, the dependence of entropy, H-bonds and hydrophobic interactions on temperature and ionic strength are taken into consideration as described in AGADIR2-1s.[37,38]

Assumptions

We have opted for a two-window sampling approximation, which assumes that the probability of finding more than two ordered segments in the same polypeptide chain is too low to be considered (the simple one window will deviate too much from reality for peptides with >50 residues). Further, we assume there is no energetic coupling between the two windows. Finally, we do not consider aggregation intermediates. This means that we consider aggregates as a single molecular species or structural state in competition with the folded protein, the β-turn and α-helical conformations. One simplification in this approximation is that we do not consider β-hairpins as aggregating motifs. The main reason for this shortcoming is the absence of a satisfactory algorithm to predict β-hairpin stability.

Acknowledgements

This work was supported, in part, by EU grant QLRI-CT2-2002-00241. J.W.H.S. and F.R. were supported by International Prize Traveling Fellowships from the Wellcome Trust. Thanks to Sara Quirk for reading this manuscript. We are grateful to Lars Juhl Jensen for the human proteome set.

References

1. Dobson, C. M. (2002). Protein-misfolding diseases: getting out of shape. *Nature*, **418**, 729–730.
2. Chiti, F., Taddei, N., Baroni, F., Capanni, C., Stefani, M., Ramponi, G. & Dobson, C. M. (2002). Kinetic partitioning of protein folding and aggregation. *Nature Struct. Biol.* **9**, 137–143.
3. Chiti, F., Stefani, M., Taddei, N., Ramponi, G. & Dobson, C. M. (2003). Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature*, **424**, 805–808.
4. Blake, C. C., Serpell, L. C., Sunde, M., Sandgren, O. & Lundgren, E. (1996). A molecular model of the amyloid fibril. *Ciba Found. Symp.* **199**, 6–15, 15–21, 40–46.
5. Serpell, L. C., Sunde, M. & Blake, C. C. (1997). The molecular basis of amyloidosis. *Cell Mol. Life Sci.* **53**, 871–887.
6. Blake, C. & Serpell, L. (1996). Synchrotron X-ray studies suggest that the core of the transthyretin amyloid fibril is a continuous beta-sheet helix. *Structure*, **4**, 989–998.
7. Serpell, L. C., Blake, C. C. & Fraser, P. E. (2000). Molecular structure of a fibrillar Alzheimer's A beta fragment. *Biochemistry*, **39**, 13269–13275.
8. Lopez De La Paz, M., Goldie, K., Zurdo, J., Lacroix, E., Dobson, C. M., Hoenger, A. & Serrano, L. (2002). *De novo* designed peptide-based amyloid fibrils. *Proc. Natl Acad Sci. USA*, **99**, 16052–16057.
9. Dobson, C. M. (2001). The structural basis of protein folding and its links with human disease. *Phil. Trans. R. Soc. ser. B*, **356**, 133–145.
10. Escamilla-Fernandez, A. M., Rousseau, F., Schymkowitz, J. W. & Serrano, L. (2004). Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. In the press.
11. Guerois, R., Nielsen, J. E. & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* **320**, 369–387.
12. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradovic, Z. (2002). Intrinsic disorder and protein function. *Biochemistry*, **41**, 6573–6582.
13. Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S. *et al*. (2001). Intrinsically disordered protein. *J. Mol. Graph. Model.* **19**, 26–59.
14. Williams, R. M., Obradovi, Z., Mathura, V., Braun, W., Garner, E. C., Young, J. *et al*. (2001). The protein non-folding problem: amino acid determinants of intrinsic order and disorder. *Pac. Symp. Biocomput.* 2001, pp. 89–100.
15. Dyson, H. J. & Wright, P. E. (2002). Coupling of folding and binding for unstructured proteins. *Curr. Opin Struct. Biol.* **12**, 54–60.
16. Tompa, P. (2002). Intrinsically unstructured proteins. *Trends Biochem. Sci.* **27**, 527–533.
17. de la Paz, M. L. & Serrano, L. (2004). Sequence determinants of amyloid fibril formation. *Proc. Natl Acad Sci. USA*, **101**, 87–92.
18. Lo Conte, L., Ailey, B., Hubbard, T. J., Brenner, S. E., Murzin, A. G. & Chothia, C. (2000). SCOP: a structural classification of proteins database. *Nucl. Acids Res.* **28**, 257–259.
19. Uversky, V. N. (2002). Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* **11**, 739–756.
20. Linding, R., Jensen, L. J., Diella, F., Bork, P.,

Gibson, T. J. & Russell, R. B. (2003). Protein disorder prediction: implications for structural proteomics. *Structure (Cambridge)*, **11**, 1453–1459.

21. Linding, R., Russell, R. B., Neduva, V. & Gibson, T. J. (2003). GlobPlot: exploring protein sequences for globularity and disorder. *Nucl. Acids Res.* **31**, 3701–3708.
22. Dill, K. A. (1997). From Levinthal to pathways to funnels. *Nature Struct. Biol.* **4**, 10–19.
23. Onuchic, J. N., Socci, N. D., Luthey-Schulten, Z. & Wolynes, P. G. (1996). Protein folding funnels: the nature of the transition state ensemble. *Fold. Des.* **1**, 441–450.
24. Dobson, C. M. (1999). Protein misfolding, evolution and disease. *Trends Biochem. Sci.* **24**, 329–332.
25. Bucciantini, M., Calloni, G., Chiti, F., Formigli, L., Nosi, D., Dobson, C. M. & Stefani, M. (2004). Prefibrillar amyloid protein aggregates share common features of cytotoxicity. *J. Biol. Chem.* 2004;.
26. Demeester, N., Mertens, C., Caster, H., Goethals, M., Vandekerckhove, J., Rosseneu, M. & Labeur, C. (2001). Comparison of the aggregation properties, secondary structure and apoptotic effects of wild-type, Flemish and Dutch N-terminally truncated amyloid beta peptides. *Eur. J. Neurosci.* **13**, 2015–2024.
27. Kayed, R., Head, E., Thompson, J. L., McIntire, T. M., Milton, S. C., Cotman, C. W. & Glabe, C. G. (2003). Common structure of soluble amyloid oligomers implies common mechanism of pathogenesis. *Science*, **300**, 486–489.
28. Si, K., Lindquist, S. & Kandel, E. R. (2003). A neuronal isoform of the aplysia CPEB has prion-like properties. *Cell*, **115**, 879–891.
29. Bucciantini, M., Giannoni, E., Chiti, F., Baroni, F., Formigli, L. Zurdo, J. *et al*. (2002). Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature*, **416**, 507–511.
30. Ventura, S., Lacroix, E. & Serrano, L. (2002). Insights into the origin of the tendency of the PI3-SH3 domain to form amyloid fibrils. *J. Mol. Biol.* **322**, 1147–1158.
31. Viguera, A. R., Jimenez, M. A., Rico, M. & Serrano, L. (1996). Conformational analysis of peptides corresponding to beta-hairpins and a beta-sheet that represent the entire sequence of the alpha-spectrin SH3 domain. *J. Mol. Biol.* **255**, 507–521.
32. Ventura, S., Zurdo, J., Narayanan, S., Parreno, M., Mangues, R., Reif, B. *et al*. (2004). Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case. *Proc. Natl Acad Sci. USA*, **101**, 7258–7263.
33. Li, W., Jaroszewski, L. & Godzik, A. (2002). Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics*, **18**, 77–82.
34. Gasteiger, E., Jung, E. & Bairoch, A. (2001). SWISS-PROT: connecting biomolecular knowledge *via* a protein database. *Curr. Issues Mol. Biol.* **3**, 47–55.
35. Chandonia, J. M., Walker, N. S., Lo Conte, L., Koehl, P., Levitt, M. & Brenner, S. E. (2002). ASTRAL compendium enhancements. *Nucl. Acids Res.* **30**, 260–263.
36. Promponas, V. J., Enright, A. J., Tsoka, S., Kreil, D. P., Leroy, C. Hamodrakas, S. *et al*. (2000). CAST: an iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts. *Bioinformatics*, **16**, 915–922.
37. Munoz, V. & Serrano, L. (1995). Elucidating the folding problem of helical peptides using empirical parameters. II. Helix macrodipole effects and rational modification of the helical content of natural peptides. *J. Mol. Biol.* **245**, 275–296.
38. Munoz, V. & Serrano, L. (1997). Development of the multiple sequence approximation within the AGADIR model of alpha-helix formation: comparison with Zimm–Bragg and Lifson–Roig formalisms. *Biopolymers*, **41**, 495–509.
39. Hooft, R. W., Sander, C., Scharf, M. & Vriend, G. (1996). The PDBFINDER database: a summary of PDB, DSSP and HSSP information with added value. *Comput. Appl. Biosci.* **12**, 525–529.

Edited by A. R. Fersht

(Received 1 March 2004; received in revised form 21 June 2004; accepted 22 June 2004)