Aggregation of natively folded proteins: a theoretical approach

2007 J. Phys.: Condens. Matter 19 285221

(http://iopscience.iop.org/0953-8984/19/28/285221)

IOP PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **19** (2007) 285221 (10pp) doi:10.1088/0953-8984/19/28/285221

Aggregation of natively folded proteins: a theoretical approach

Antonio Trovato[1,2], **Amos Maritan**[1,2,3] **and Flavio Seno**[1,2,3]

[1] CNISM, Unità di Padova, Via Marzolo 8, 35131 Padova, Italy
[2] Dipartimento di Fisica 'G. Galilei', Università di Padova, Via Marzolo 8, 35131 Padova, Italy
[3] Sezione INFN, Dipartimento di Fisica, Università di Padova, Italy

Received 5 October 2006, in final form 12 January 2007
Published 25 June 2007
Online at stacks.iop.org/JPhysCM/19/285221

Abstract

The reliable identification of β-aggregating stretches in protein sequences is essential for the development of therapeutic agents for Alzheimer's and Parkinson's diseases, as well as other pathological conditions associated with protein deposition. While the list of aggregation related diseases is growing, it has also been shown that many proteins that are normally well behaved can be induced to aggregate *in vitro*. This fact suggests the existence of a unified framework that could explain both folding and aggregation. By assuming this universal behaviour, we have recently introduced an algorithm (PASTA: prediction of amyloid structure aggregation), which is based on a sequence-specific energy function derived from the propensity of two residue types to be found paired in neighbouring strands within β-sheets in globular proteins. The algorithm is able to predict the most aggregation-prone portions of several proteins initially unfolded, in excellent agreement with experimental results. Here, we apply the method to a set of proteins which are known to aggregate, but which are natively folded. The quality of the prediction is again very high, corroborating the hypothesis that the amyloid structure is stabilized by the same physico-chemical determinants as those operating in folded proteins.

(Some figures in this article are in colour only in the electronic version)

1. Introduction

Biological macromolecules such as proteins, lipids and nucleic acids have the ability to assemble into functional complexes in a highly regulated manner within densely crowded environments [1, 2]. Moreover, the balance between normal and pathological self-association has been carefully tuned by molecular evolution [3, 4]. Failures of the regulatory mechanisms do, however, occur and may result in conditions such as Alzheimer's and Creutzfeldt–Jakob diseases and type II diabetes. Such diseases are associated with the deposition in tissue of

pathogenic aggregates that are composed largely of misfolded proteins in the form of amyloid fibrils or plaques [5, 6].

Recent work suggests that amyloid aggregates can also be induced *in vitro* with proteins not associated with known deposition diseases [3, 6, 7]. Such observations have led to the suggestion that the ability to form amyloid fibrils may be a common characteristic of polypeptide chains [3, 6], although individual propensities vary greatly with the sequence and the environmental conditions. This view has recently been supported by theoretical arguments [8–10] that have shown that simple considerations of geometry and symmetry are sufficient to explain, within the same sequence-independent framework, the emergence of a limited menu of native-like conformations for a single chain and β-aggregate for multiple chains.

Images obtained with transmission electron microscopy or atomic force microscopy reveal that the fibrils usually consist of typically two to six protofilaments, each about 2–5 nm in diameter [11]. These protofilaments generally twist together to form fibrils that are typically 7–13 nm wide [12] or associate laterally to form long ribbons that are 2–5 nm high and up to 30 nm wide. X-ray fibre diffraction data have shown that the protein or peptide molecules are arranged so that the polypeptide chain forms β-strands that run perpendicular to the long axis of the fibril [11].

Solid-state nuclear magnetic resonance, x-ray micro- or nano-crystallography and other techniques such as systematic protein engineering coupled to site-directed spin or fluorescence labelling have transformed our ability to gain insight into structures of fibrillar aggregates with residue-specific detail [13–16]. These studies have allowed the identification of regions of the sequence that form and stabilize the cross-β of the fibrils. One frequent characteristic emerging from these studies is the parallel in-register arrangement of β-strands in the fibril core [13, 14].

At the same time, several theoretical approaches have been developed to predict aggregation-prone regions in the amino-acid sequence of a full-length protein [17–21]. All such approaches focus on predicting the β-aggregation propensity of a sequence stretch by itself. From another perspective, in [22] a sequence-specific energy function has been derived from the propensity of two residue types to be found paired in neighbouring strands within β-sheets in globular proteins. This function was implemented within an algorithm (PASTA: prediction of amyloid structure aggregation) that was able to show that parallel in-register arrangement of sequence portions participating in the fibril cross-β core is favoured in most cases. It also predicted the most aggregation-prone portions of an initially unfolded polypeptide chain, in excellent agreement with experimental observations.

In this paper we want to extend the application of PASTA by showing its ability to also predict aggregation for natively structured proteins that are known to aggregate under particular physiological conditions. The quality of the results that we present confirm again the validity of this approach, which is based on just a statistical analysis of globular proteins. This corroborates the hypothesis that amyloids and the native state of globular proteins are stabilized by the same physico-chemical determinants.

2. Methods

The algorithm PASTA is based on an energy function for specific β-aggregation which can be obtained by an ensemble of globular proteins, including all-α, all-β, α/β and $\alpha + \beta$ proteins. For this work we have used the top500H database, selected by means of resolution (1.8 Å or better), low sequence homology (less than 30%) and other criteria of quality (such as low clash score [23], few atoms whose main-chain bond angles deviate too much from typical geometries, no unusual amino acids with main-chain substitutions, no free-atom

refinements) [24]. Hydrogens were added to the crystallography structures in standardized geometry [23].

All the possible occurring instances n_{ab} of a given ab residue pair, where both a and b run over the 20 amino acids types, are partitioned in four different classes:

(1) n^p_{ab} = number of times the two residues are facing each other on neighbouring parallel β-strands;
(2) n^a_{ab} = number of times the two residues are facing each other on neighbouring anti-parallel β-strands;
(3) n^c_{ab} = number of times the distance between the C_α of the two residues is less than 6.5 Å, without participating in ordered β-geometry (generic contacts);
(4) n^d_{ab} = all the other cases in which the residues are not in contact.

We include in the counting all pairs except those formed by consecutive residues along the chain. The participation to β-strand is assigned by using the DSSP algorithm [25].

We assume that the database is a system in thermodynamic equilibrium at a single temperature that is assumed to be roughly constant for all proteins in the data-bank, even though each single sequence is a separate system. The propensity $p_{ab}(x)$ of the pair to be found in one of the four pairing types x is then given by $p_{ab}(x) = \exp(-E^x_{ab})$, where the Es are effective adimensional energies, since the Boltzmann factor $\kappa_B T$ is absorbed in their definition [26].

Propensities are defined as the ratio of the observed frequencies over the expected probabilities, which is in turn estimated as the frequency observed over all pairs. They tell whether there is more ($E^x_{ab} < 0$) or less ($E^x_{ab} > 0$) probability of finding the residue pair ab in the pairing type x with respect to what is found in the reference state built by considering all possible residue pairs. In this way, propensities are not biased by the fact that, for instance, parallel β-pairing is less common than anti-parallel β-pairing in the top500 dataset. We can estimate the effective energies through the relations:

$$E^p_{ab} = -\log\left(\frac{\frac{n^p_{ab}}{n_{ab}}}{\frac{\sum_{a'b'} n^p_{a'b'}}{\sum_{a'b'} n_{a'b'}}}\right) \tag{1}$$

$$E^a_{ab} = -\log\left(\frac{\frac{n^a_{ab}}{n_{ab}}}{\frac{\sum_{a'b'} n^a_{a'b'}}{\sum_{a'b'} n_{a'b'}}}\right) \tag{2}$$

and similarly for E^c_{ab} and E^d_{ab}. Since the numbers n^x_{ab} can be very small or even zero (e.g. for Proline and Cysteine), the large statistical error due to small statistics is dramatically amplified by the logarithm in equations (1) and (2). We therefore used an averaging procedure [18] so that, for example, $E^p_{ab} = (E^{p+}_{ab} + E^{p-}_{ab})/2$, where E^{p+}_{ab}, E^{p-}_{ab} are the energies obtained from equations (1) and (2) when adding or subtracting a single event to the observed number of cases (whenever $n^x_{ab} < 2$, 0.5 is used in place of $n^x_{ab} - 1$). The resulting estimate of E^p_{ab} is a better representative of the confidence window.

We want to predict the specific aggregation pattern of a pair of identical proteins composed by N amino-acids $\{a_k\}_{1\leqslant k\leqslant n}$ as determined by the specific β-pairing (either parallel or anti-parallel) of the sequence stretch of length L beginning at position i on the first chain, with the sequence stretch of the same length, beginning on position j on the second chain. We assume that only a single stretch per sequence participates in the β-pairing and that all other residues are not involved and are found in a disordered non-compact conformation. We assume further that the energies E^d_{ab} of all pairs involving these latter residues can be neglected, since $n^d_{ab} \sim 0$ and $E^d_{ab} \simeq 0$. It has been verified [22] that the results do not change upon the inclusion of non-contacting pair terms.

The overall energy for a given parallel or anti-parallel pattern is then given by:

$$\epsilon^{p}_{i,j}(L) = \sum_{k=0}^{L-1} E^{p}_{a^{1}_{i+k},a^{2}_{j+k}} - L\Delta s \tag{3}$$

$$\epsilon^{a}_{i,j}(L) = \sum_{k=0}^{L-1} E^{a}_{a^{1}_{i+k},a^{2}_{j+L-1-k}} - L\Delta s \tag{4}$$

where the over-scripts 1 and 2 correspond to the first and second chains, respectively, and $S = L\Delta s$ is the total entropy loss due to the β-ordering of the L residues pairs, with Δs corresponding to the entropy loss per residue pair. In the calculations performed in this paper we use $\Delta s = -0.2$.

Since we can associate an energy to all possible β-pairing conformations, it is convenient to introduce a partition function:

$$Z = \sum_{i,j,L\geqslant 4} [\exp(-\lambda\epsilon^{p}_{i,j}(L)) + \exp(-\lambda\epsilon^{a}_{i,j}(L))] \tag{5}$$

where we use $\lambda = 2.0$ as an adimensional factor setting the energy scale. It has been observed [22] that parameters Δs and λ need not to be fine tuned and can be changed within a 20% range without affecting the final results. Since protein structures were assumed to be obtained in equilibrium at the same physiological temperature, $\lambda = 2.0$ implies that the energy unit used in this work is roughly 1.2 kcal mol^{-1}, with $\Delta s = 0.24$ kcal mol^{-1}. Such values should be taken with proper caution, due to the many approximations involved in the standard derivation of statistical potentials and to the bias inherent with the choice of the reference state.

In order to visualize the results it is convenient to introduce a probability $p(k)$ which tells us to which extent a residue at position k is more likely to aggregate into an ordered β structure with respect to other ones. $p(k)$ is given by:

$$p(k) = \frac{\sum_{i,j,L\geqslant 4}(\alpha_{i,L,k} + \alpha_{j,L,k}/2L)[\exp(-\lambda\epsilon^{p}_{i,j}) + \exp(-\lambda\epsilon^{a}_{i,j})]}{Z} \tag{6}$$

where $\alpha_{i,L,k} = 1$ if residue k belong to the stretch of length L going from i to $i + L - 1$ and 0 otherwise.

Similarly it is possible to obtain the two dimensional probability $p_2(k, m)$ of two residues to be found paired to each other within an ordered β-structure through the relation:

$$p_2(k,m) = \frac{\sum_{i,j,L\geqslant 4}(\alpha_{i,L,k}\alpha_{j,L,k}/L)[\delta_{k-m+j-i}\exp(-\lambda\epsilon^{p}_{i,j}) + \delta_{k+m+1-L-j-i}\exp(-\lambda\epsilon^{a}_{i,j})]}{Z} \tag{7}$$

where k and m label the residues in two different chains and $\delta_\gamma = 1$ if $\gamma = 0$ and 0 otherwise. Based on $p_2(h, m)$, a β-pairing contact map can be produced where the orientation (parallel or anti-parallel) and the register of the best pairings can easily be traced out.

3. Results

In [22] the PASTA algorithm was applied on natively unfolded proteins, because the method employs values of the intrinsic propensity of residues pairs to aggregate and does not take into account the presence and type of secondary and tertiary structures in the analysed polypeptide chain. Indeed, it is well known [6] that the presence of structure in the initial non-aggregated state of the protein is an important determinant of aggregation. Therefore, here we want to

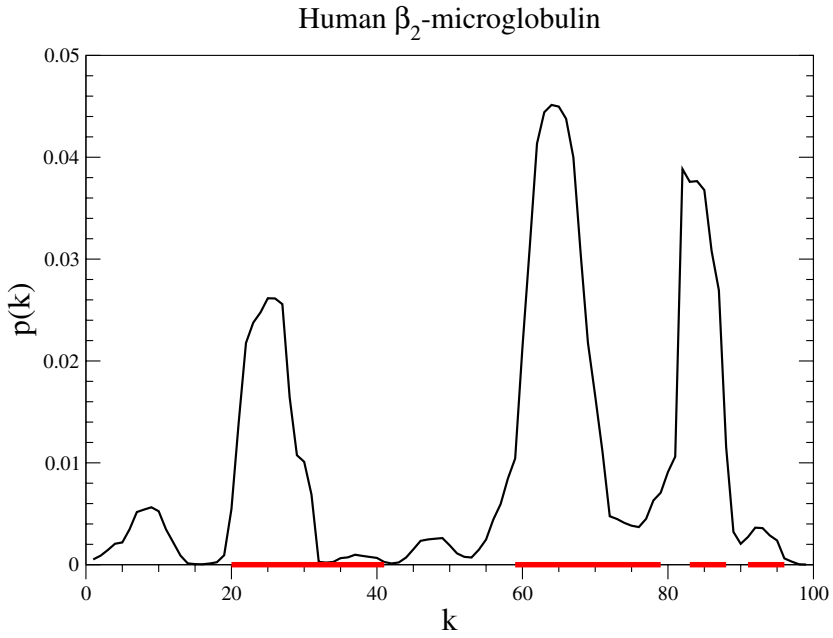

Figure 1. Plot of amyloid propensity $p(k)$ (equation (6)) for human β_2-microglobulin. The sequence fragments that are known to form amyloid fibrils in isolation [32, 33] are represented by thick lines (red online) along the k axis.

verify whether PASTA can also be predictive in this case, even without introducing any specific term to consider the original structure of the protein. The energy functions introduced in equations (3) and (4) can be used to compare different segment lengths, but more often we will use the single residue propensity $p(k)$ defined in equation (6) to take into account other low-energy pairings that could be close competitors of the lowest-energy one or the contact map $p_2(k, m)$ defined by equation (7), which can readily visualize the register of the aggregation segments.

First, we study human β_2-microglobulin, a normally soluble protein that aggregates into pathogenic fibrils either at low pH [27] or under physiological conditions when divalent copper is present [28]. There is quite clear evidence that dialysis-related amyloidosis pathogenesis results as a destabilization of the native structure of β_2-microglobulin followed by the formation of nucleating species that eventually form amyloid fibrils [29, 30]. While a complete structural model for the fibrillar aggregate formed by the full protein is lacking, it has been found that a few sequence fragments form fibrils in isolation, namely: (a) Ser 20 to Lys 41 [31]; (b) Asp 59 to Ala 79 [32]; (c) Asn 83 to Ser 88 [33]; and (d) Lys 91 to Asp 96 [33]. These segments are shown as thick bars (red online) in figure 1, together with the aggregation profile predicted by PASTA through equation (6). In figure 2 we show the β-pairing contact map $p_2(k, m)$, where a general compendium of the general features predicted by PASTA can be found. By looking at figure 1 we notice that four out of the five most significant peaks predicted by the algorithm fall within one of the above-mentioned segments. In particular (see figure 2), we predict parallel in-register alignment for the regions 20–31, 82–88 and 92–95 and anti-parallel in-register alignment with different possible registries, 60–84, 61–68, 61–71, 60–67, where in all cases the same sequence fragment is paired with itself, as predicted by PASTA on more general grounds [22]. By inspection of figure 2, we see that the only false positive, in the

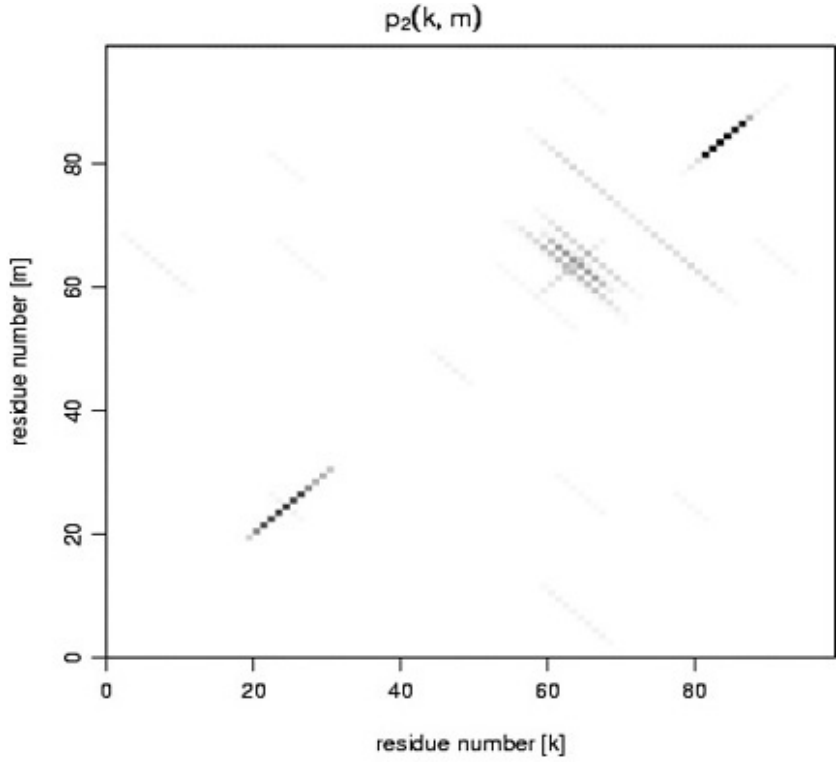

Figure 2. β-pairing contact map (equation (7)) for β_2-microglobulin. The darker the spot, the higher the probability of the corresponding β-contact. Dark lines along the main diagonal (0, 0) to (1, 1) are signals of parallel in-register aggregation. Other lines parallel to the same diagonal indicate parallel not in-register alignment. Lines along the other diagonal describe anti-parallel aggregation.

region 3–12, is prone to form a β-sheet by coupling with the region 60–69 and not with itself. The fact that we predict anti-parallel alignment in the central 60–80 region is apparently in contradiction with the theoretical prediction in [20], where however only windows of size 5 were used to implement the method. We checked that, by assuming the same constraint ($L = 5$ in equations (3) and (4)), we also obtain a parallel in-register alignment in the same region, even though the competition with anti-parallel alignments is quite close. By using $L = 6$ fragments, as done in [33] where parallel alignment is assumed as a default in their prediction method, we instead obtain a preferred anti-parallel alignment (62–67 with itself). We also observe that other theoretical methods [20, 21] implemented on this protein are not able to individuate the segment 91–96 as a possible sequence fragment prone to aggregate.

The second protein that we study is sperm whale myoglobin (for the sake of comparison with [18], when using horse myoglobin we obtain very similar results), which is a compact and highly soluble protein without any native state properties to suggest that it has a predisposition to form amyloid fibrils. Whereas the latter are characteristically rich in β-sheets, native myoglobin lacks any elements of such structure and has most of its sequence arranged in well-defined α-helices. Moreover, all partially folded states of the protein characterized so far are significantly helical [34], although, as in most proteins, there is evidence for the presence of more extended conformations in aggregates and precipitates [34]. In [35] it has been observed that myoglobin forms large quantities of fibrillar structure, at 65 °C and pH 9.0, conditions

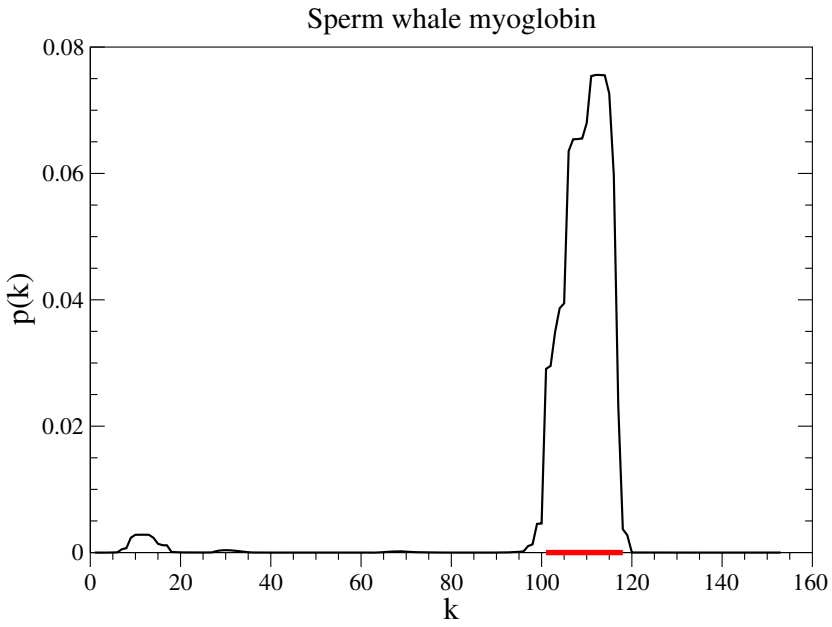

Figure 3. Plot of amyloid propensity $p(k)$ (equation (6)) for sperm whale myoglobin. The sequence fragment known to form amyloid fibrils in appropriate conditions '*in vitro*' [36] is represented by a thick line (red online) along the k axis.

under which the native fold is substantially destabilized. This was the first example of protein, not related to any disease, which was induced to aggregate *in vitro*. By applying PASTA to myoglobin (figure 3) we see a very strong signal, indicating the helical fragment 101–116 (part of what is usually termed the G helix, from residue 100 to 118) as a promoter of the fibril stability. The predicted alignment is parallel in-register. This result is consistent with recent experimental work [36] in which the G helix was found to form amyloid fibrils. The other theoretical method that was used to study this protein is TANGO [18], which also finds the G helix as a possible candidate for promoting aggregation but which also obtains two other regions with a similar propensity, namely the A and E helices, which is at variance with our method.

Human muscle acylphosphatase is a relatively simple α/β protein consisting of 98 residues that aggregates in a very well-defined manner under appropriate conditions, ultimately forming highly organized amyloid fibrils [37]. The aggregation propensity profile obtained with PASTA is shown in figure 4, together with the sequence fragments, which were shown to aggregate in isolation [37]. Our algorithm correctly localizes the stretches 34–53 and 86–98, although the latter is found with a much lower propensity than the former. It also predicts 10–25 as a possible aggregation-promoting region with a low propensity. The aggregation propensity profile predicted by PASTA is in this case similar overall to the one predicted by TANGO.

Another protein used in [18] to benchmark the performance of TANGO is French bean plastocyanin, which is a protein that consists largely of β-sheet, with reverse turns and loops between the strands of the sheet, and one short helix [38]. Seven fragments have been seen experimentally to aggregate [38], and they are reported as thick bars (red online) in figure 5 together with the propensity $p(k)$. Remarkably, the overlap between PASTA predictions and these fragments is very high, and the overall performance of the method is higher than TANGO, which is, for example, completely missing the region 24–43.

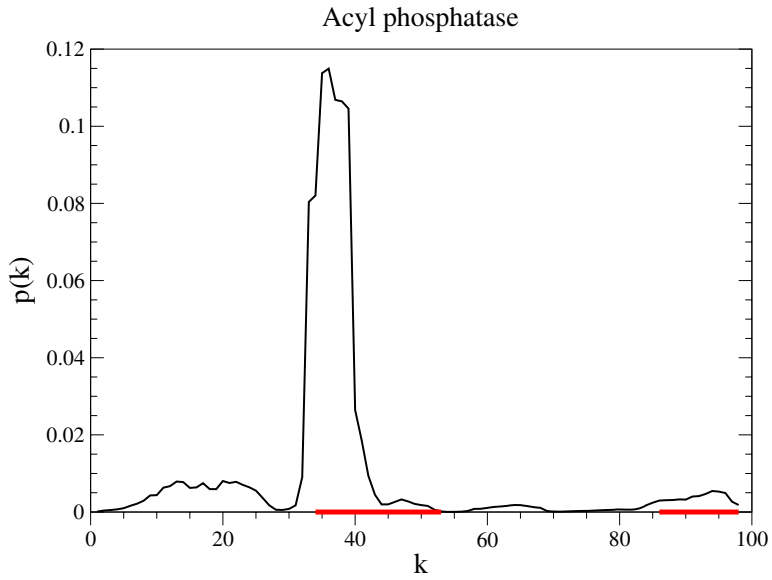

Figure 4. Plot of amyloid propensity $p(k)$ (equation (6)) for human muscle acylphosphatase. The sequence fragments known to aggregate in isolation [37] are represented by thick lines (red online) along the k axis.

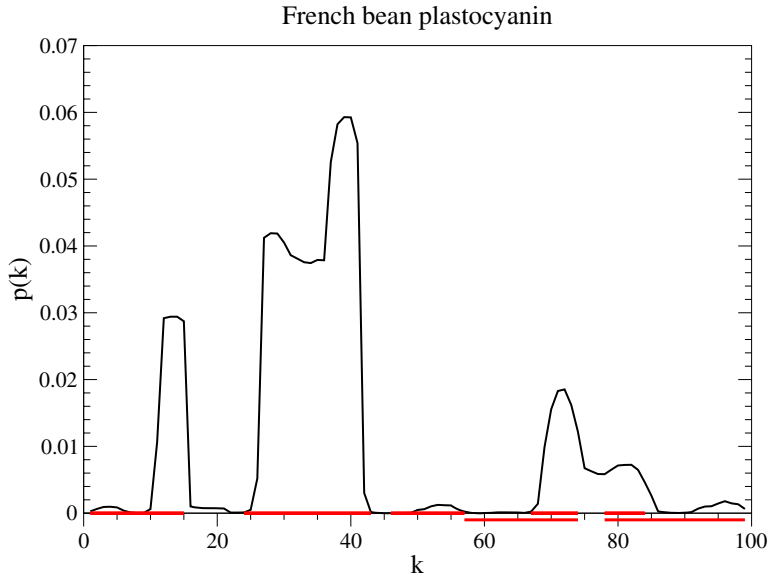

Figure 5. Plot of amyloid propensity $p(k)$ (equation (6)) for French bean plastocyanin. The sequence region fragments known to aggregate in isolation [38] are represented by thick lines (red online) along the k axis.

4. Conclusion

In this manuscript we have employed the method PASTA, which was recently introduced [22] to predict amyloidogenic sequence stretches as well as the registry of the inter-molecular

hydrogen bonds formed between them, in the case of natively unfolded proteins, to study four proteins which aggregate but which are natively folded.

The predictions obtained using the method for these proteins are consistent with experimental findings. However, they give a more complete description of the possible aggregation patterns with respect to other theoretical methods because PASTA allows for an exhaustive analysis of possible patterns that involve a fragment of a protein forming a beta-sheet with any other fragment of another protein, in either parallel or anti-parallel orientation.

This suggests that PASTA is a powerful theoretical algorithm for studying protein aggregation. The fact that the whole computational approach is derived from the knowledge of globular proteins underscores the universality of the physico-chemical mechanisms underlying amyloid fibril formation. Moreover, it indicates that the structure and stabilizing interactions existing in the apparently monotonous amyloid or amyloid-like fibrils are of the same essential nature as those determining structural and functional diversity in globular proteins.

Acknowledgments

We thank Fabrizio Chiti for several illuminating discussions. This work was supported by Programmi di Rilevanza Scientifica di Rilevante Interesse Nazionale no. 2005027330 in 2005.

References

[1] Ellis R J 2001 *Trends Biochem. Sci.* **26** 597–604
[2] Minton A P 2000 *Curr. Opin. Struct. Biol.* **10** 34–9
[3] Dobson C M 1999 *Trends. Biochem. Sci.* **24** 329–32
[4] Vendruscolo M, Zurdo J, MacPhee C E and Dobson C M 2001 *Phil. Trans. R. Soc.* B **3456** 133–45
[5] Selkoe D J 2003 *Nature* **426** 900–4
[6] Chiti F and Dobson C M 2006 *Annu. Rev. Biochem.* **75** 333–6
[7] Chiti F, Webster P, Taddei N, Clark A, Stefani M, Ramponi G and Dobson C M 1999 *Proc. Natl Acad. Sci. USA* **96** 3590–4
[8] Banavar J R, Hoang T X, Maritan A, Seno F and Trovato A 2004 *Phys. Rev.* E **70** 041905
[9] Hoang T X, Trovato A, Seno F, Banavar J R and Maritan A 2004 *Proc. Natl Acad. Sci. USA* **101** 7960–4
[10] Hoang T X, Marsella L, Trovato A, Seno F, Banavar J R and Maritan A 2006 *Proc. Natl Acad. Sci. USA* **103** 6883–8
[11] Sunde M and Blake C 1997 *Adv. Protein Chem.* **50** 123–59
[12] Serpell L C *et al* 2000 *J. Mol. Biol.* **300** 1033–9
[13] Petkova A T *et al* 2002 *Proc. Natl Acad. Sci. USA* **99** 16742–7
[14] Torok M, Milton S, Kayed R, Wu P and McIntire T 2002 *J. Biol. Chem.* **278** 37530–5
[15] Der-Sarkissian A, Jao C C, Chen J and Langen R 2003 *J. Biol. Chem.* **278** 37530–5
[16] Margittai M and Langen R 2004 *Proc. Natl Acad. Sci. USA* **102** 15871–6
[17] Yoon S and Welsh S J 2004 *Protein Sci.* **13** 2149–60
[18] Fernandez-Escamilla A M, Rosseau F, Schymkowitz J and Serrano L 2004 *Nat. Biotechnol.* **350** 379–92
[19] Pawar A P *et al* 2005 *J. Mol. Biol.* **350** 379–92
[20] Tartaglia G G, Cavalli A, Pallarin R and Caflisch A 2005 *Protein Sci.* **14** 2723–34
[21] Galzitskaya O V, Garbuzynskiy S O and Lobanov M Y 2006 *PLoS Comput. Biol.* **2** e177
[22] Trovato A, Chiti F, Maritan A and Seno F 2006 *PLoS Comput. Biol.* **2** e170
[23] Word J M *et al* 1999 *J. Mol. Biol.* **285** 1711–33
[24] Lovell S C *et al* 2003 *Proteins* **50** 437–50
[25] Kabsch W and Sander C 1983 *Biopolymers* **22** 2577–637
[26] Samudrala R and Moult J 1998 *J. Mol. Biol.* **275** 895–916
[27] MacParland V J, Kalverda A P, Brown A, Kirwin-Jones P, Hunter M G, Sunde M and Radford S E 2000 *Biochemistry* **39** 8735–46
[28] Eakin C M, Knight J D, Morgan C J, Gelfand M A and Miranaker A D 2002 *Biochemistry* **41** 10646–56
[29] Chiti F *et al* 2001 *J. Mol. Biol.* **307** 379–91
[30] McParland V J, Kalverda A P, Homans S W and Radford S E 2002 *Nat. Struct. Biol.* **9** 326–31

[31] Kozhukh G V, Hagihara Y, Kawakami T, Hasegawa K, Naiki H and Goto Y 2002 *J. Biol. Chem.* **277** 1310–5
[32] Jones S, Manning J, Kad N M and Radford S E 2003 *J. Mol. Biol.* **325** 249–57
[33] Ivanova M I, Thompson M J and Eisenberg D 2006 *Proc. Natl Acad. Sci. USA* **103** 4079–82
[34] Eliezer D, Yao J, Dyson H J and Wright P E 1998 *Nat. Struct. Biol.* **5** 148–55
[35] Fandrich M, Fletcher M A and Dobson C M 2001 *Nature* **410** 165–6
[36] Fandrich M, Forge V, Buder K, Karlis M, Dobson C M and Diekman S 2003 *Proc. Natl Acad. Sci. USA* **100** 15463–8
[37] Chiti F, Taddei N, Baroni F, Capanni C, Stefani M, Ramponi G and Dobson C M 2002 *Nat. Struct. Biol.* **9** 137–43
[38] Dyson H J, Sayre J R, Merutka G, Shin H C, Lerner R A and Wright P E 1992 *J. Mol. Biol.* **226** 819–35