Universitat Autònoma de Barcelona
Departament de Bioquímica i Biologia Molecular

Sequential and structural determinants of protein aggregation

Natalia Sánchez de Groot

January 2010

PART I:
Protein aggregation *in vitro*

Amyloid fibril formation by bovine cytochrome *c*

Natalia S. de Groot and Salvador Ventura [*]

Departament de Bioquimica i Biologia Molecular, Universitat Autonoma de Barcelona and Institut de Biotecnologia i de Biomedicina, 08193 Bellaterra (Barcelona), Spain

Abstract. Bovine heart cytochrome *c* is an all-α globular protein containing a covalently bound heme group. Prolonged incubation at 75°C in mild alkaline solution damages the prosthetic group and results in permanent unfolding of the polypeptide chain. Under this conditions, cytochrome *c* aggregates into fibrillar structures. Characterization by transmission electron microscopy and thioflavin-T binding assays shows that these species posses the characteristics of fibrils associated with the family of amyloid diseases. Our findings indicate that destabilization of the native fold of this highly α-helical protein can lead to its polymerization into β-sheet rich structures and suggest that this process does not depend on the population of partially folded monomeric states with extensive β-sheet structure.

Keywords: Amyloid formation, cytochrome *c*, protein misfolding, protein denaturation, helical proteins

Abbreviations: CD = circular dichroism; FTIR = Fourier-transform infrared.

1. Introduction

The deposition of amyloid fibrils has been linked to a variety of slow-onset degenerative diseases, such as Alzheimer's disease, senile systemic amyloidosis, Parkinson's disease, dyalisis-related amyloidosis, and transmissible spongiform encephalopathies [1–4]. The proteins responsible for these diseases do not share structural or sequential identities [5]. In spite of this diversity, all amyloid fibrils display similar structural features, exhibiting a cross-β structure. In the last few years, proteins unrelated to any known human disease have been found to convert *in vitro* into higher order structures that also present a cross-β structure and fulfill all characteristics of amyloid fibrils [6–12]. This has suggested that amyloid represents a generic form of polypeptide conformation, and most peptides/proteins have the potential to form amyloid-like structures under appropriate conditions [13].

The mechanism of fibril assembly is still controversial. For long time it has been accepted that amyloid fibril formation involved the docking of monomeric partially folded states, which display at least partial β-sheet structure [8,14–19]. Nevertheless, recent studies suggest that whereas this can occur in specific cases it may not be the general rule. This way, it has been shown that in the case of myoglobin, an ordinary all-α protein, amyloid fibril formation correlates whit environments in which the protein backbone is unfolded, rather than with conditions that may allow population of partially structured states [11,20]. In this context, it is likely that the study of different protein models with predominant helical secondary structure should provide new insights into the onset of the aggregation process.

[*] Corresponding author. Tel.: +34 93 581 41 47; Fax: +34 93 581 12 64; E-mail: salvador.ventura@uab.es.

Bovine heart cytochrome c is a small heme protein with 104 amino acid residues, and is an important electron-transfer protein in the respiratory chain. The three-dimensional native structure of this protein fold has been well characterized [21,22], and consists of four α-helices forming a compact core around the covalently attached heme moiety without any β-sheet segment. In this study we show that destabilization of the native fold of this helical protein promotes the formation of amyloid fibrils from an essentially unfolded state.

2. Material and methods

Heart bovine cytochrome c, thioflavin T (ThT), and Trizma base were purchased from Sigma. Unless otherwise mentioned, all solutions were made in 50 mM Tris-HCl buffer (pH 9.0). Controlled heating of protein samples were obtained by using an Erycom PCR system for the desired incubation time.

Circular dichroism. Circular dichroism (CD) spectra in the far- and near-UV region were obtained by using a Jasco 710 spectropolarimeter at 25°C. Protein was assayed at 5–100 μM. Ten accumulations were averaged to obtain each spectrum.

Dye binding assays. Thioflavin-T binding assays were carried out using aliquots of 50 μl drawn from 90 μM protein samples incubated as indicated above. These aliquots were stained with 0.5% Thioflavin-T, washed twice with H_2O and air-dyed. Samples were viewed under UV light using a Leica fluorescence microscope.

Transmission electron microscopy. Samples containing 90 μM protein were incubated as indicated above. A 5 μl aliquot was then placed on carbon-coated copper grids, and allow them to stand for 2 min. The grids were then washed with distilled water and stained with 2% uranyl acetate for another 2 min prior to analysis using a Hitachi H-7000 transmission electron microscope operating at accelerating voltages of 75 kV.

Fourier-transform infrared (FTIR) spectroscopy analysis. Aggregates were dried for 1 h in a speed-vac system prior to analysis to reduce H_2O interference in the infrared spectra. The structure of the dry aggregates was directly analysed in a Bruker Tensor FT-IR spectrometer. FT-IR spectrum of the native protein was acquired after air-drying the protein solution. For each spectrum, 20 interferograms were collected and averaged. All processing procedures were carried out so as to optimize the quality of the spectrum in the amide I region, between 1550 and 1700 cm^{-1}. Second derivatives of the amide I band spectra were used to determine the frequencies at which the different spectral components were located.

3. Results and discussion

Bovine heart cytochrome c posses a predominant helical secondary structure (48.2%) under mild alkaline conditions (pH 9.0) at room temperature (calculated using the Contin method with CDPro suite[1]). This is illustrated by the far-UV CD spectrum shown in Fig. 1, which displays the typical 210 and 222 nm minima. Heating of cytochrome c to 75°C results in partial unfolding of the protein and reduced helical content (38.8%), as denoted by the decrease in the signal strength of the minima at 222 nm and a shift of the band at 210 nm toward lower values (Fig. 1). When the protein is heated just for 5 min

[1] CDPro suite at http://lamar.colostate.edu/~sreeram/CDPro/main.html.

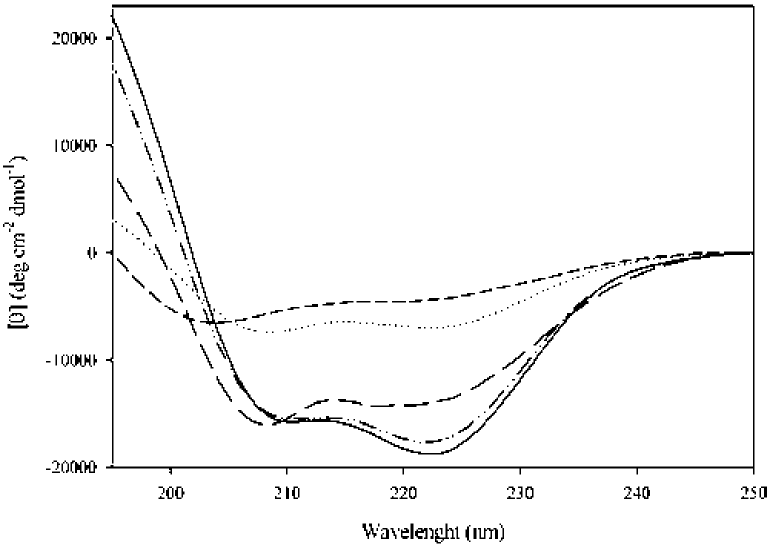


Fig. 1. Far UV-CD spectra of bovine heart cytochrome *c* samples at 25°C (—), at 75°C (– –), incubated at 75°C for 5 min and cooled down to 25°C (–··–), incubated at 75°C for 4 h and cooled down to 25°C (· · · ·) and incubated at 75°C for 12 h and cooled down to 25°C (- - - -). Buffer was Tris-HCl 50 mM pH 9.0.

these structural changes are almost fully reversible and the original CD spectra shape and helical content (46.5%) are recovered upon cooling (Fig. 1). Incubation of the protein at 75°C for 4 h results in permanent conformational changes with a significant decrease in helical structure (21.4%) and large increase in random coil conformation (Fig. 1). Further heating of the sample up to 12 h promotes permanent protein unfolding and almost complete loss of the helical content (0.1%). Under these conditions the polypeptide chain is found mainly in random coil conformation (Fig. 1).

Prolonged heating of the protein at 75°C in 50 mM Tris-HCl pH 9.0 also causes a significant loss of the typical red colour present in cytochrome *c* solutions. No protein or heme group aggregation was detected after 12 h incubation at 75°C and 5 μM protein concentration. Furthermore, no soluble dissociated prosthetic group could be found upon gel filtration of the protein solution, being all colour associated to the protein fraction (data not shown). The visible absorption spectra of a bovine cytochrome *c* preparation incubated at 25°C for 12 h shows the typical Soret-band maximum at 408 nm due to the heme iron in its oxidized form (Fig. 2A). In addition, bands at 219 and 550 nm attributable to the presence of some reduced cytochrome *c* species, are also detected (Fig. 2B). The same cytochrome *c* solution heated at 75°C for 12 h lacks any reduced-state associated band (Fig. 2B) and exhibits a 6 fold decrease in absorbance at 408 nm (Fig. 2A). These results suggest irreversible structural changes in the covalently bound heme group of cytochrome *c* upon prolonged incubation at 75°C. These changes in the prosthetic group appear to promote unfolding of the polypeptide chain, as assessed by CD. The extent of protein conformational change depends on the time of incubation and presumably on the degree of heme group alteration. Our data are consistent with the notion that the heme group in cytochrome *c* is not only the redox center of the protein, but is also critical for maintaining the native structure: its removal produces apocytochrome *c*, and has been shown to cause disruption of the native fold and loss of most of the secondary structure under physiological conditions in different cytochromes [23–28].

Recently it has been reported that prolonged incubation at room temperature of an apo form of cytochrome c_{552} from *Hydogenobacter thermophilus* resulted in the formation of protein aggregates with

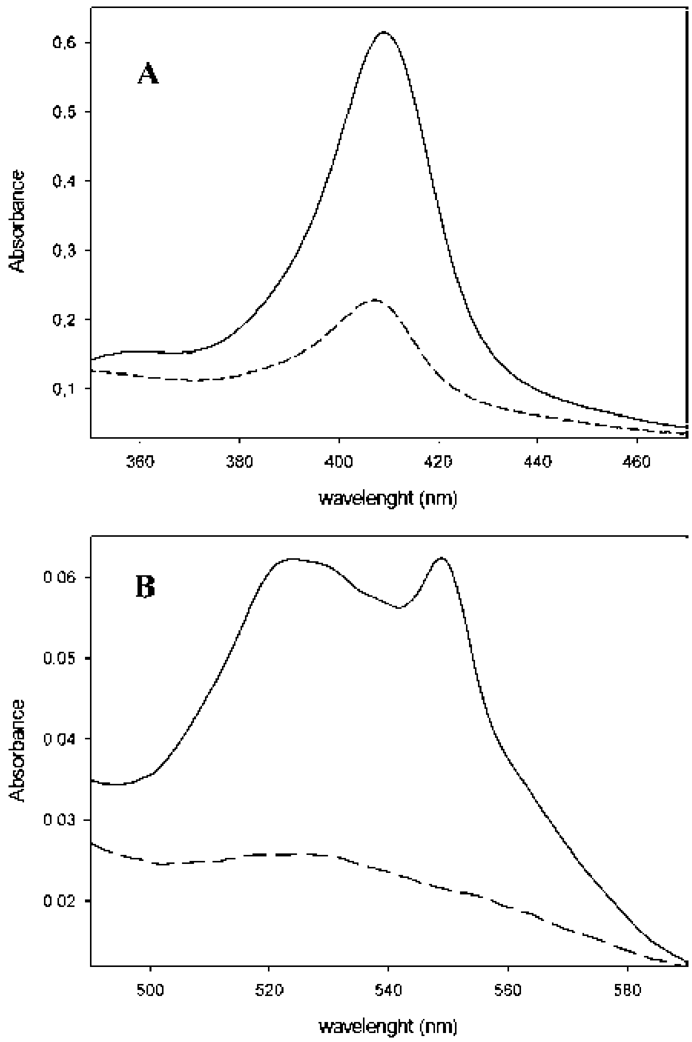


Fig. 2. Absorption spectra of 5 μM bovine heart cytochrome *c* solutions after incubation for 12 h at 25°C (—), or 75°C (--). The band at 408 nm correspond to oxidized heme forms (A) and the bands at 519 and 550 nnm to reduced species (B).

amyloid-like properties [28]. In addition, it has been shown that incubation of the apo form of muscle myoglobin, another helical heme protein, at pH 9.0 and 65°C causes the formation of large quantities of fibrillar structures [20]. Under these conditions the native fold of apomyoglobin is, as it happens with apocytochrome c_{552}, substantially destabilized [20,28]. The conformational properties of bovine heart cytochrome *c* when incubated at 75°C in a pH 9.0 solution for 12 h resemble very much those exhibited by the apo forms of the above mentioned and related heme proteins. Hence, we focused on the possibility that it may also aggregate into amyloid-like structures. We screened for conditions that might promote protein aggregation of bovine cytochrome *c* and found that aggregation was strongly dependent on protein concentration (data not shown). Slight precipitation of cytochrome *c* was detected at 90 μM protein concentration at the end of the 12 h incubation period at 75°C. Further characterization of the protein aggregates by electron microscopy revealed the presence of fibrillar structures, which resemble those formed by disease-related proteins (Fig. 3A). Thioflavin-T (Th-T) is an amyloid azo-free diag-

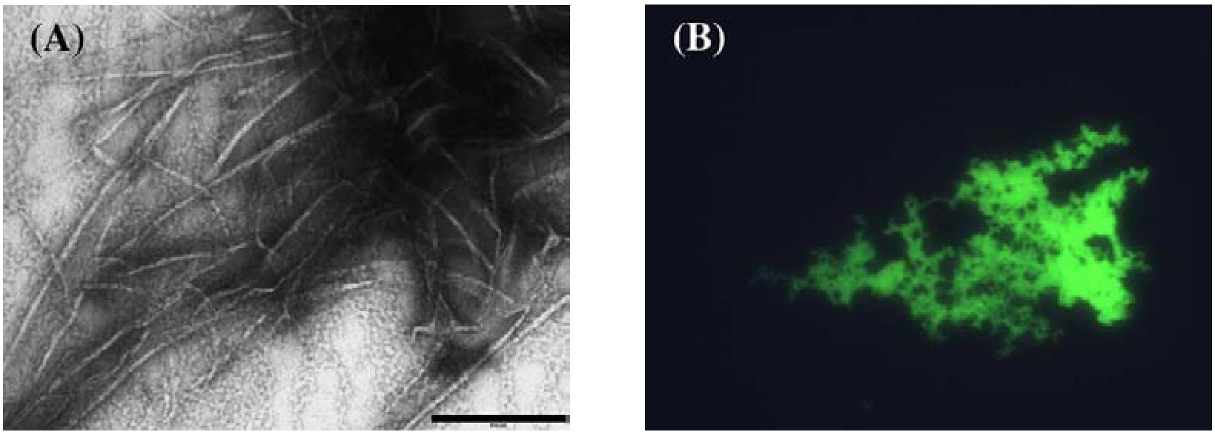


Fig. 3. Features of the fibrils formed by bovine heart cytochrome c after 12 h incubation at 75°C and pH 9.0 at 90 μM protein concentration. (A) Transmission electron micrograph of negatively stained aggregated protein (bar $=$ 500 nm). (B) Thioflavin-T fluorescence of stained amyloid-like material.

nostic dye that by a so far unknown mechanism specifically interacts with the crossed-β-pleated sheet structure common to a variety of amyloid fibrils. Binding of this dye to cytochrome c fibrils was probed by fluorescence microscopy. The areas rich in protein fibrous material appeared stained with Th-T, giving a bright green–yellow fluorescence against a dark background (Fig. 3B), reinforcing the amyloid-like nature of the aggregated protein.

To further characterize the nature of the cytochrome c aggregates we used FTIR spectroscopy. In myoglobin the formation of amyloid-like aggregates resulted in a significant reduction of the α-helix content and new formation of β-sheet structure [20]. To see whether this was also the case for cytochrome c, we recorded the FTIR spectrum of the native and aggregated states in the amide I region. The difference spectrum between both states shows a strong formation of new β-sheet structure with a concomitant loss of α-helix content upon aggregation (Fig. 4). Because native cytochrome c posses only α-helical secondary structure and because amyloid fibrils are always associated with β-sheet structure, the aggregated β-sheets are constructed from residues that form α-helices in the folded protein. Thus, it is clear that to enable fibril formation, these structural elements need to be previously unfolded.

Taken together our data argue that long-lasting incubation of bovine cytochrome c at 75°C in a pH 9.0 solution somehow damages/denatures the heme group in the native protein, resulting in a loss of the cooperative native structure in these conditions. This leads to unfolding of the protein, which backbone adopts chiefly a random coil conformation. Consequently, the polypeptide chain becomes exposed to solvent allowing the establishment of intermolecular interactions, resulting in concentration dependent protein aggregation. At low protein concentration this aggregation occurs in the form of ordered amyloid fibrils of the type formed by disease related proteins. The behaviour of bovine heart cytochrome c is especially interesting because, as it happens with cytochrome c_{552} and muscle myoglobin [28,29], is a helical protein devoid of β-sheet elements in the native state, whereas amyloid fibrils posses mainly β-sheet structure. Besides, predictions of the secondary structure content of bovine cytochrome c clearly show the absence of stretches with β-sheet propensity for this protein sequence (43.5% α-helix and 0% β-sheet are predicted with the PSIPRED algorithm[2]). Thus, bovine cytochrome c constitutes yet another example in which secondary structure propensity and amyloid fibril formation are not related.

[2] PSIPRED at http://www.psipred.net.

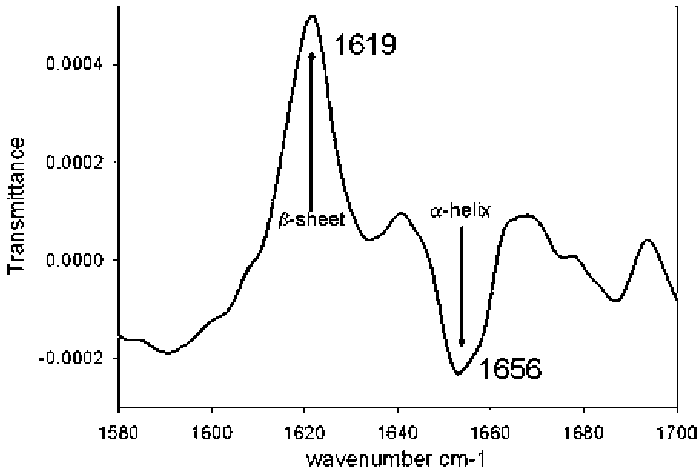


Fig. 4. Difference FTIR spectrum in the amide I region between the aggregated and native states of bovine heart cytochrome *c*. Arrows indicate the increase of β-sheet and decrease of α-helix structures upon aggregation.

The present results support the idea that amyloid fibril formation is an intrinsic property of many polypeptide chains with independence of the conformation of their native state [4]. Although the structural, thermodynamic and kinetic factors determining the polymerisation of helical cytochrome *c* into β-sheet rich fibrils should be studied in much more detail before we can understand the rules underlying this self-assembling process, our data suggest that, resembling what happens to myoglobin, amyloid fibril formation occurs for this protein under conditions in which the polypeptide backbone is predominantly unfolded. Thus, it appears that amyloid fibril formation by bovine cytochrome *c* does not require significant population of partially folded intermediates with β-sheet conformation as those reported for other protein models. This behaviour is not exclusive of highly helical proteins, since it is now clear that amyloid fibrils can be formed by very short peptides [29,30] or polyaminoacids [31,32], which neither fold nor populate partially structured states. Hence, it is likely that the presence of unfolded protein regions may be a general requirement for the formation of amyloid fibrils.

The data reported herein show that an intact prosthetic group permits the recovery of the native cytochrome *c* structure after a moderate conformational stress situation, avoiding prolonged exposition of unfolded protein regions to solvent and thus reducing aggregation propensity. This observation provides a possible explanation for the role of covalently linked heme groups in this protein family and supports the suggestion that natural protein sequences have evolved in part to code for structural characteristics other than those included in the native fold, such us avoidance of aggregation.

Acknowledgements

S.V. is supported by a “Ramón y Cajal” project awarded by the MCYT and co-financed by the Universitat Autonoma de Barcelona.

References

[1] S.Y. Tan and M.B. Pepys, *Histopathology* **25** (1994), 403–414.
[2] J.W. Kelly, *Curr. Opin. Struct. Biol.* **8** (1998), 101–106.

[3] J.C. Rochet and P.T. Lansbury, Jr., *Curr. Opin. Struct. Biol.* **10** (2000), 60–68.

[4] C.M. Dobson, *Phil. Trans. R. Soc. Lond.* **356** (2001), 133–146.

[5] M. Sunde and C.C. Blake, *Quart. Rev. Biophys.* **31** (1998), 1–39.

[6] J.I. Guijarro, M. Sunde, J.A. Jones, I.D. Campbell and C.M. Dobson, *Proc. Natl. Acad. Sci. USA* **95** (1998), 4224–4228.

[7] S.V. Litvinovich, S.A. Brew, S. Aota, S.K. Akiyama, C. Haudenschild and K.C. Ingham, *J. Mol. Biol.* **280** (1998), 245–258.

[8] F. Chiti, P. Webster, N. Taddei, A. Clark, M. Stefani, G. Ramponi and C.M. Dobson, *Proc. Natl. Acad. Sci. USA* **96** (1999), 3590–3594.

[9] M. Ramirez-Alvarado, J.S. Merkel and L. Regan, *Proc. Natl. Acad. Sci. USA* **97** (2000), 8979–8984.

[10] V. Villegas, J. Zurdo, V.V. Filimonov, F.X. Aviles, C.M. Dobson and L. Serrano, *Protein Sci.* **9** (2000), 1700–1708.

[11] M. Fandrich, M.A. Fletcher and C. Dobson, *Nature* **410** (2001), 165–166.

[12] R.J. Ellis and T.J. Pinheiro, *Nature* **416** (2002), 483–484.

[13] C.M. Dobson, *Trends in Biochemical Sciences* **24** (1999), 329–332.

[14] J.W. Kelly, *Proc. Natl. Acad. Sci. USA* **95** (1998), 930–932.

[15] Z. Lai, W. Colon and J.W. Kelly, *Biochemistry* **35** (1996), 6470–6482.

[16] D.R. Booth, M. Sunde, V. Belloti, C.V. Robinson, W.L. Hutchinson, P.E. Fraser, P.N. Hawkins, C.M. Dobson, S.E. Radford, C.C. Blake and M.B. Pepys, *Nature* **385** (1997), 787–793.

[17] V.N. Uversky, J. Li and A.L. Fink, *J. Biol. Chem.* **276** (2001), 10737–10744.

[18] V.J. McParland, N.M. Kad, A.P. Kalverda, A. Brown, P. Kirwin Jones, M.G. Hunter, M. Sunde and S.E. Radford, *Biochemistry* **39** (2000), 8735–8746.

[19] R. Khurana, J.R. Gilleppie, A. Talapatra, L.J. Minert, C. Jonesiu-Zanetti, I. Millett and A.L. Fink, *Biochemistry* **40** (2001), 3525–3535.

[20] M. Fandrich, V. Forge, K. Buder, M. Kittler, C.M. Dobson and S. Diekmann, *Proc. Natl. Acad. Sci. USA* **100** (2003), 15463–15468.

[21] P.X. Qi, D.L. Di Stefano and A.J. Wand, *Biochemistry* **33** (1994), 64082–6417.

[22] G.W. Bushnell, G.V. Louie and G.D. Brayer, *J. Mol. Biol.* **214** (1990), 585–595.

[23] C.D. Moore and J.T.J. Lecomte, *Biochemistry* **29** (1990), 1984–1989.

[24] Y. Feng and S.G. Sligar, *Biochemistry* **30** (1991), 10150–10155.

[25] M.F. Jeng, S.W. Englander, G.A. Elove, A.J. Wand and H. Roder, *Biochemistry* **29** (1990), 10433–10437.

[26] D. Hamada, M. Hoshino, M. Kataoka, A.L. Fink and Y. Goto, *Biochemistry* **32** (1993), 10351–10358.

[27] Q. Feng, S.G. Sligar and A.J. Wand, *Nat. Struct. Biol.* **1** (1994), 30–35.

[28] T.A. Pertinhez, M. Bouchard, E.J. Tomlinson, R. Wain, S.J. Ferguson, C.M. Dobson and L.J. Smith, *FEBS Let.* **495** (2001), 184–186.

[29] E. Gazit, *Curr. Med. Chem.* **9** (2002), 1725–1735.

[30] M. Lopez De La Paz, K. Goldie, J. Zurdo, E. Lacroix, C.M. Dobson, A. Hoenger and L. Serrano, *Proc. Natl. Acad. Sci. USA* **99** (2002), 16052–16057.

[31] M. Fandrich and C.M. Dobson, *EMBO J.* **21** (2002), 5682–5690.

[32] M.F. Perutz, B.J. Pope, D. Owen, E.E. Wanker and E. Scherzinger, *Proc. Natl. Acad. Sci. USA* **99** (2002), 5596–5600.

PART II:
Protein sequence and aggregation prediction

the FEBS Journal

Mutagenesis of the central hydrophobic cluster in Aβ42 Alzheimer's peptide

Side-chain properties correlate with aggregation propensities

Natalia Sánchez de Groot, Francesc X. Aviles, Josep Vendrell and Salvador Ventura

Departament de Bioquímica i Biologia Molecular and Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona

Keywords

protein aggregation; protein misfolding; Alzheimer's disease; green fluorescent protein; *Escherichia coli*

Correspondence

S. Ventura, Departament de Bioquímica i Biologia Molecular and Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, E-08193 Bellaterra, Spain

Fax: +34 935811264

Tel: +34 935814147

E-mail: salvador.ventura@uab.es

(Received 14 October 2005, revised 7 December 2005, accepted 14 December 2005)

doi:10.1111/j.1742-4658.2005.05102.x

Protein misfolding and deposition underlie an increasing number of debilitating human disorders. Alzheimer's disease is pathologically characterized by the presence of numerous insoluble amyloid plaques in the brain, composed primarily of the 42 amino acid human β-amyloid peptide (Aβ42). Disease-linked mutations in Aβ42 occur in or near a central hydrophobic cluster comprising residues 17–21. We exploited the ability of green fluorescent protein to act as a reporter of the aggregation of upstream fused Aβ42 variants to characterize the effects of a large set of single-point mutations at the central position of this hydrophobic sequence as well as substitutions linked to early onset of the disease located in or close to this region. The aggregational properties of the different protein variants clearly correlated with changes in the intrinsic physicochemical properties of the side chains at the point of mutation. Reduction in hydrophobicity and beta-sheet propensity resulted in an increase of *in vivo* fluorescence indicating disruption of aggregation, as confirmed by the *in vitro* analysis of synthetic Aβ42 variants. The results confirm the key role played by the central hydrophobic stretch on Aβ42 deposition and support the hypothesis that sequence tunes the aggregation propensities of polypeptides.

More than 20 different diseases including Alzheimer's disease (AD), spongiform encephalopathies, type II diabetes mellitus and Parkinson's disease are associated with the occurrence of protein aggregates called amyloid fibrils [1–5]. Alzheimer's disease is a progressive neurodegenerative disorder characterized by the patient's memory loss and impairment of cognitive abilities that affects a substantial fraction of the elderly [6]. The extracellular amyloid is found both at neuropil sites and in blood vessel walls in the brain and is widely believed to be involved in the progressive neurodegeneration of the disease [7]. The principal component of these lesions is a hydrophobic 40–43 amino acid peptide [8] called β-amyloid peptide (Aβ). The most abundant forms found in amyloid plaques are a 40-mer (Aβ40) and a longer isoform containing two C-terminally additional hydrophobic amino acids (Aβ42). Although less abundant, Aβ42 is more amyloidogenic than Aβ40 and is the major component of neuritic plaques [9,10]. Aβ is produced from a much larger protein termed the amyloid precursor protein (APP) as a cleavage product of secretases whose enzymatic components are suggested to include presenilins and β-site APP cleaving enzyme [11]. Most mutations associated with early onset familial AD occur in APP and presenilins [12–15]. Interestingly, such mutations are also associated with increased production of Aβ42 [12–15]. The overexpression of structurally normal APP that results from an extra gene in trisomy 21 (Down syndrome) almost invariably leads to the premature occurrence of classic AD neuropathology during middle adult years [16]. Together, these findings provide strong evidence for the role of Aβ42 in AD and AD-like pathology.

Abbreviations

AD, Alzheimer's disease; Aβ, β-amyloid peptide; Aβ42, 42-amino acid human β-amyloid peptide; APP, amyloid precursor protein; CHC, central hydrophobic cluster; FAD, familial Alzheimer's disease; GFP, green fluorescent protein; PIMT, protein isoaspartate methyltransferase; Th-T, Thioflavin T; WT, wild-type.

Aβ42 contains a central hydrophobic cluster (CHC) (Leu17-Val18-Phe19-Phe20-Ala21) that has been suggested to be important for peptide aggregation. In this way, the substitution of two or more hydrophobic amino acid residues between positions 17 and 20 in a synthetic Aβ peptide encompassing residues 10–43 results in increased solubility [17]. Replacement of single residues in this region with proline also decreases the aggregation propensity of a peptide comprising residues 15–23 [18]. A short seven-residue fragment, KLVFFAE, is able to form ordered amyloid fibrils and, more interestingly, LVAFF and derived peptides have been shown to bind to Aβ42 and act as potent inhibitors of amyloid formation [19,20]. The CHC does not only influence the rate of Aβ monomer assembly into fibrils but itself appears to be part of the β-sheet core of the mature fibrils [21,22]. Among CHC residues, position 19 has been shown to strongly affect the folding, assembly and fibril structure of Aβ [18,23,24], thus being an excellent target to test effects of sequence changes on Aβ42 peptide aggregation propensity.

The Aβ42 peptide is difficult to synthesize, purify and study because of its very low solubility in physiological buffers. This property has impeded the analysis of large sets of synthetic variants in order to understand the sequential determinants of Aβ42. However, several indirect *in vivo* methods have been developed recently that are able to monitor the aggregation of very insoluble polypeptides by connecting an easily monitored function in a reporter protein to the aggregation propensity of the fused polypeptide. Waldo *et al.* demonstrated that the fusion of the green fluorescent protein (GFP) to insoluble proteins dramatically reduces its folding ability in *E. coli*, showing that GFP can be used as a reporter for the folding of upstream fusion proteins [25]. Also, Hecht and coworkers fused Aβ42 to GFP and exploited the system to isolate variants with reduced aggregation propensity from a randomly generated library [26]. We have used this system to analyse the effects of mutation of the central amino acid in the CHC of Aβ42 on the aggregation propensity of the peptide and compared the results thus obtained with the behaviour of relevant synthetic Aβ42 peptides. We have also tested the system's potential to foresee the depositional properties of Aβ42 mutants related to familial AD. Overall, we find that the aggregational propensities of the different variants can be correlated with the characteristics of the changed residues, allowing deduction of those side chain properties related to Aβ42 aggregation in this particular *in vivo* system.

Results

Expression, solubility and fluorescence of Phe19 mutants in Aβ42-GFP fusions

The adopted approach, originally developed by Hecht *et al.* [26], uses the wild-type (WT) Aβ42 gene inserted as a fusion protein upstream of the GFP sequence and under the control of the T7 promoter, with the two sequences separated by a 12-residue linker. *E. coli* cells transformed with this vector express a high amount of Aβ42–GFP fusion but exhibit little fluorescence, indicating that the presence of the aggregation-prone Aβ42 peptide strongly interferes with the development of the GFP native structure and thus with the emission of fluorescence, as previously reported [26].

To elucidate if the identity of the residue in the central position of the CHC of Aβ42 influences its deposition in this particular system, we systematically substituted Phe19 in the Aβ42-GFP fusion by the rest of 19 natural amino acids using PCR, generating a collection of 20 different vectors differing only in the residue at position 19 of Aβ42 that were used to transform *E. coli* cells. Three hours after induction of protein expression, cultured cells were collected, incubated at 4 °C overnight to ensure equilibrium, and their emitted fluorescence analysed. As expected, the intensity of the green fluorescence varied from clone to clone (Fig. 1A). The dynamic range of fluorescence comparing the most fluorescent mutant (F19D) to the less fluorescent mutant (F19I), including the WT sequence, was approximately five fold (Fig. 1B).

To confirm that the different levels of fluorescence exhibited by the mutants were not simply related to different protein expression levels in *E. coli*, the amount of Aβ42–GFP fusions in induced whole cell extracts was monitored by SDS/PAGE. All clones expressed the fusion protein at comparable levels (see Supplementary material). Thus, differences in fluorescence can be attributed to variations in the proportion of active GFP from clone to clone, since fluorescence indicates both a correct tertiary folding of the GFP moiety and proper chromophore maturation. These processes have been shown to occur relatively slowly inside the cells and, consequently, the presence of fused aggregation-prone sequences, such as Aβ42, can affect GFP fluorescence emission strongly by promoting aggregation. The formation of refractile inclusion bodies was observed in transformed and induced *E. coli* cells (data not shown), suggesting that the aggregated Aβ42–GFP protein fusions accumulate into such structures, which, in fact, have been shown to share some structural features with amyloids [27]. The higher the

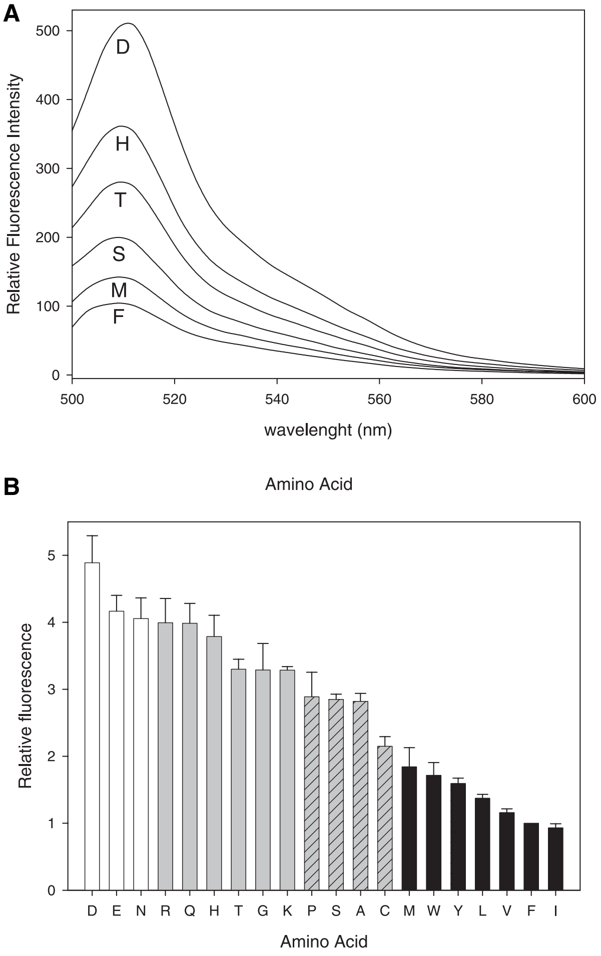


Fig. 1. Fluorescence emission by *E. coli* cells expressing wild-type (WT) and Phe19 Aβ42 mutants fused to green fluorescent protein (GFP). (A) Fluorescence spectra of selected clones. Amino acid in position 19 is indicated. (B) Fluorescence data of all Phe19 Aβ42 mutants relative to that of WT. The data are ordered by decreasing relative fluorescence at 510 nm. The bars indicate clones exhibiting < 2 (black bars), 2–3 times (grey shaded bars), 3–4 times (grey bars) and > 4 times (white bars) fluorescence increase.

aggregation propensity of the fusion protein, the lower its fluorescence emission and *vice versa*, as aggregation competes with the formation of a correctly folded GFP structure. Then, it follows that substitutions in position 19 of Aβ42 significantly affect its aggregation propensity.

Amyloidogenic properties of WT and F19D Aβ42 peptides

The results shown above refer to different aggregation propensities of Aβ42 mutants when fused to GFP and analysed inside *E. coli*. It has been previously shown that the data obtained in this system mirror the effects of identical changes on synthetic Aβ42 peptides [26]. To assess that this also applies to this study, the aggregation properties of WT Aβ42 were compared to those of the mutant exhibiting the highest fluorescence *in vivo* (F19D) using two 42-mer peptides obtained by solid phase synthesis.

Secondary structure

The secondary structure content of freshly and aged solutions of WT and F19D Aβ42 peptides were analysed using CD spectroscopy. The CD spectra of freshly dissolved peptides show that both of them are mostly in random coil conformation under the conditions of the assay (Fig. 2A). The spectrum of F19D changes little with aging, whereas a dramatic increase in β-sheet content is observed in the WT solution, as deduced from the strong CD minima at 217–220 nm. The predominant β-sheet structure found in the WT form upon aging is coincident with the described in the literature for Aβ42 amyloid fibrils or precursors [28], whereas the absence of the β-sheet signature in F19D Aβ42 spectra indicates that it is unable to assemble into such structures.

Binding to amyloid-specific dyes

The presence of polypeptidic chains in a crossed β-pleated sheet conformation is a testable characteristic of amyloid fibrils. Binding of Thioflavin T (Th-T) to amyloid fibrils induces a large increase in the fluorescence of Th-T relative to free dye [29]. Figure 2B shows the fluorescence spectra of Th-T incubated in the presence of aged WT Aβ42 or F19D peptides. While the mutant peptide exhibited little binding, a sixfold increase in the fluorescence emission maximum of Th-T occurred after binding to the WT form. Congo red, a second amyloid diagnostic dye that has also been suggested to bind to most amyloids [28] exhibits an absorbance maximum at 490 nm that shifts to red once it binds to amyloid material. Figure 2C shows the absorption spectra of Congo red incubated in the presence of aged WT Aβ42 or F19D peptides. While little Congo red binding was detected for the mutant peptide, the presence of the WT form promoted a strong increase in absorbance and a red shift of the maximum from 490 to 505 nm.

Electron microscopy

Although binding to amyloid-specific dyes has been usually attributed to the presence of amyloid fibrils,

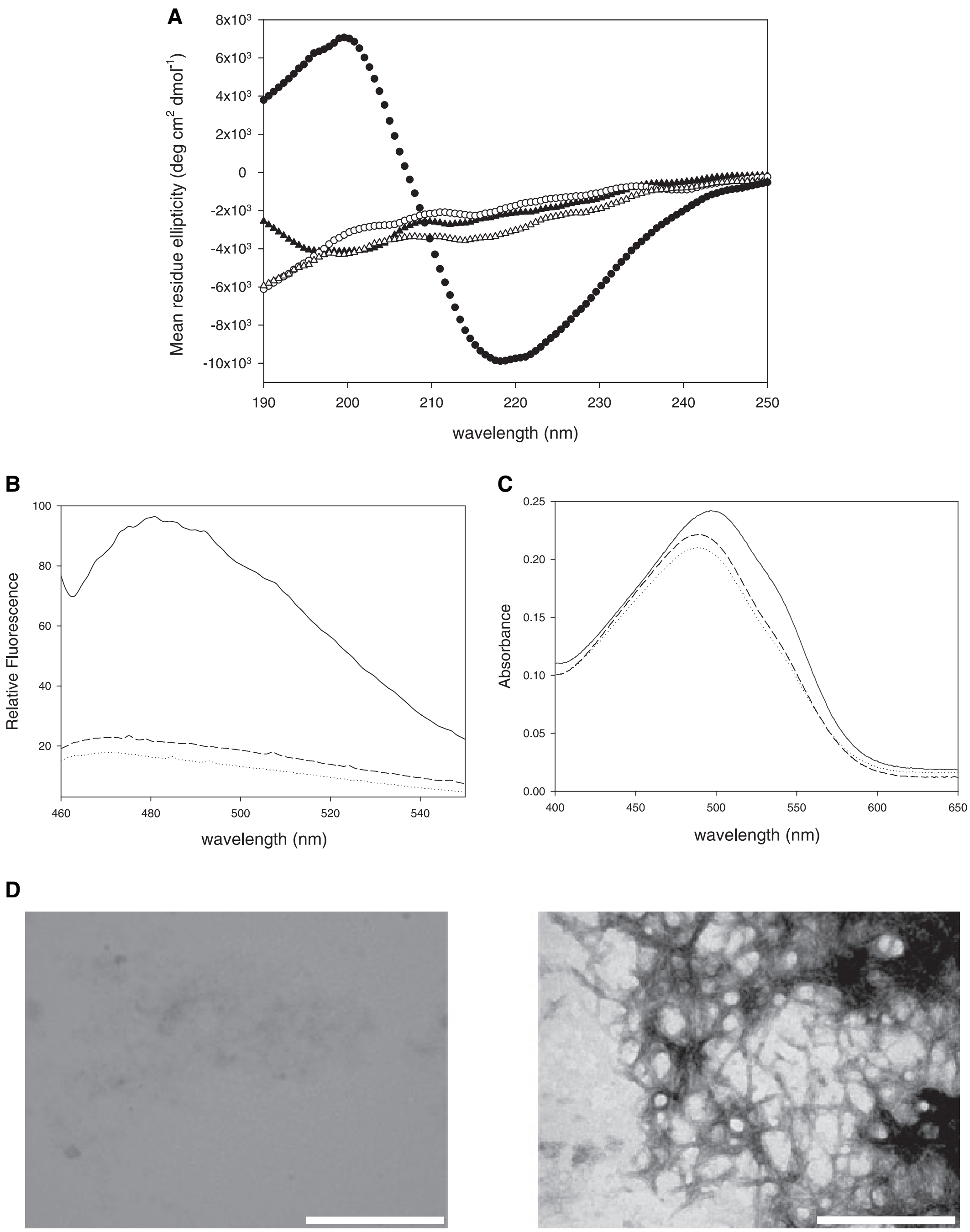


Fig. 2. Secondary structure and amyloid properties of synthetic peptides of Aβ42. (A) CD spectra in the Far-UV region of freshly (empty symbols) and aged (filled symbols) solutions of WT (circles) and Phe19Asp (triangles) Aβ42 peptides. (B) Binding of aged solutions of WT (solid line) and Phe19Asp (dashed line) Aβ42 synthetic peptides to Th-T. Thioflavin-T alone is shown as a dotted line. (C) Binding of aged solutions of WT (solid line) and Phe19Asp (dashed line) Aβ42 synthetic peptides to Congo red. Congo red alone is shown as a dotted line. (D) Representative electron microscopy images of aged solutions of Aβ42 synthetic peptides. Wild-type peptide (right) and Phe19Asp mutant peptide (left).

other protein aggregates have been shown to bind them [30,31]. Electron microscopy of aged solutions of WT and F19D peptides detected no depositions or particles for the F19D mutant whereas numerous fibrils were observed in wild-type Aβ42 samples (Fig. 2D).

Overall, the analysis of the amyloidogenic properties of these two extreme synthetic Aβ42 peptides validate the data obtained *in vivo* for the GFP fusions, suggesting that the differences in fluorescence emission might reflect different amyloid capabilities.

Correlation between fluorescence emission and side chain properties

To elucidate the basic rules underlying the observed differences in aggregation propensities we studied the correlation between the aggregation resulting from single amino acid substitutions at position 19 and the changes in the intrinsic properties of the polypeptide.

Hydrophobicity

Hydrophobic interactions have long been suggested to play an important role in protein aggregation [32]. We calculated the change in the hydrophobicity of the polypeptide chain resulting from mutation (see Experimental procedures). When the changes in hydrophobicity were plotted against the observed changes in fluorescence emission of the different Aβ42–GFP fusions, a significant correlation was detected, independent of the scale used (Fig. 3A and B).

Propensity to form β-sheet

Despite their origin, all protein aggregates are characterized by an increase of the β-sheet content respect the native conformation [32]. The propensity of a sequence to form β-sheet has been thus related to the ability of a sequence to form aggregates. When the quantified effects of the mutations on Aβ42 β-sheet propensity are plotted against the observed changes in aggregation, the correlation is found to be statistically significant despite the scatter in the plot (Fig. 3C).

Charge

Changes in the net charge of polypeptides have been shown to influence aggregation rates [33,34]. The low number of mutations implying a change in charge prevented us from obtaining significant correlations in our study. Nevertheless, charged residues rank among the most fluorescent substitutions. It is worthwhile to mention that acidic residues perform better than basic ones. This effect has been also reported for C-terminal mutants of Aβ42 peptide [35] and can be explained by analysing the effect of mutation in the net charge of the polypeptide. Aβ42 has six negative residues and three positive ones, with a net charge of

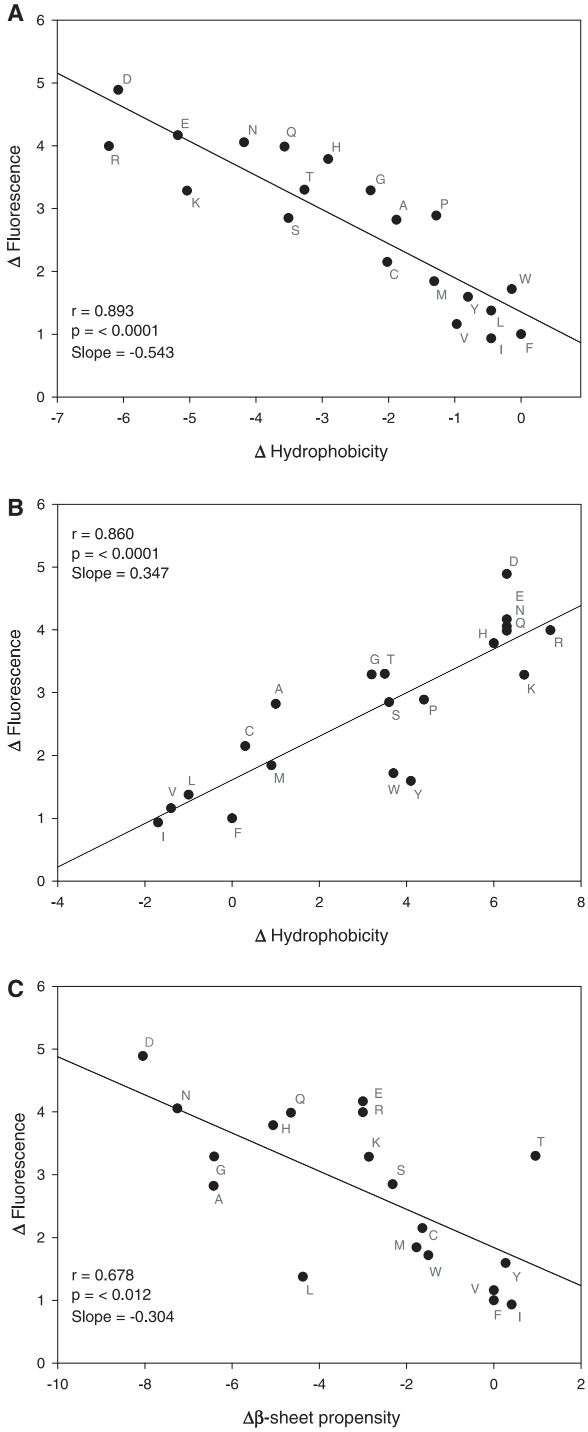


Fig. 3. Dependence of fluorescent emission on simple physicochemical properties. Change in fluorescence emission of Aβ42–GFP variants upon mutation of Phe19 plotted against: (A) the predicted change in hydrophobicity using amino acid values based on the partition coefficients from water to octanol; (B) the predicted change in hydrophobicity using amino acid values based on the hydropathicity scale from Kyte and Doolittle; (C) the predicted propensity to change from an α-helix to a β-sheet conformation.

−3. Adding one negative charge by mutating a neutral amino acid (Phe) into a Glu or Asp increases the net charge to −4, whereas mutation into a positively charged one reduces it to −2. An increase in the net charge of a polypeptide has been shown to correlate with a reduced aggregation tendency while a decrease favours aggregation [33,34], allowing us to explain the superiority of negatively charged residues over positively charged ones in reducing aggregation in Aβ42 peptide.

Prediction of fluorescence emission upon mutation

Dobson and coworkers have shown using an *in vitro* approach that hydrophobicity, β-sheet propensity and charge are independent and additive factors that can be combined in a function to predict the effect of a mutation on the aggregation rates of an unfolded polypeptide (Eqn 1). We plotted the predicted changes in aggregation rates upon mutation of position 19 according to Eqn 1 against the observed changes in fluorescence emission of the different variants when fused to GFP. The observed correlation is highly significant ($r = 0.945$; $P \leq 0.0001$) and better than that obtained from intrinsic properties alone (Fig. 4). Thus, the equation appears to be accurate in the prediction of aggregation tendencies from the changes in intrinsic polypeptide properties introduced upon mutation in this *in vivo* system, allowing for an at least qualitative prediction of how a mutation is going to affect fluorescence emission.

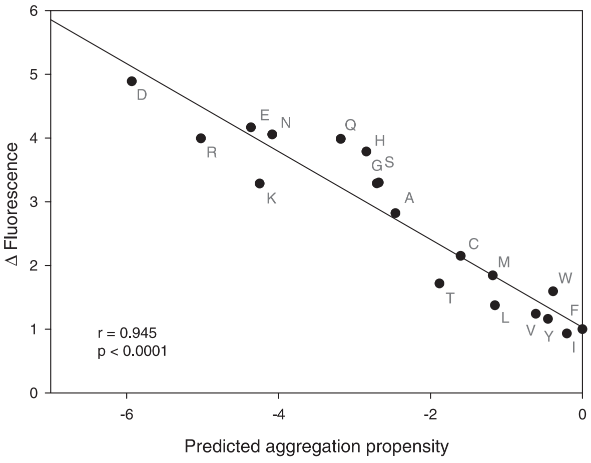


Fig. 4. Correlation of the *in vivo* emitted fluorescence with predicted aggregation propensities of Aβ42–GFP fusions. Observed changes in fluorescence emission upon mutation of Phe19 in Aβ42–GFP fusions plotted vs. the changes in aggregation propensity predicted by Eqn 1.

Table 1. Experimental fluorescence and predicted aggregation rates of Aβ42 mutations associated to familiar Alzheimer's disease.

	Mutation	OF[a]	AR[b]
Dutch	E22Q	0.67 ± 0.2	2.90 ± 0.8[c]
Arctic	E22G	0.83 ± 0.2	2.05 ± 0.3[d]
Flemish	A21G	1.77 ± 0.3	− 0.07 ± 0.3[e]

[a] Observed fluorescence, relative to that emitted by WT Aβ42–GFP fusion. [b] Aggregation rates extracted from the literature. [c] In references [40,42]. [d] In references [39,42]. [e] In references [39,42].

Mutations associated with familial AD

A set of mutations in the CHC and adjacent positions of Aβ42 is intimately associated to early onset familial AD (FAD). The substitutions include A21G, associated with a familial form of cerebral amyloid angiopathy in a Flemish kindred [36]; E22Q which causes hereditary cerebral haemorrhage with Amyloidosis-Dutch type [37] and E22G, the 'Arctic' mutation, which was linked to early onset AD in a Swedish kindred [38]. Aβ42 congeners bearing these mutations display distinct aggregation kinetics. The rate of fibril formation by the Flemish mutant is decreased relative to WT Aβ42 [39] whereas the Dutch mutant peptides aggregate substantially faster [23,29]. The Arctic peptide does not show an overall change in the rate of fibrillogenesis relative to WT Aβ, but rather accelerated protofibril formation [40]. To assess whether mutants of the Aβ42–GFP fusions would reproduce the aggregation properties reported in the literature, the effect of the Dutch, Arctic and Flemish mutations in the fluorescence emission was analysed. A decrease in fluorescence relative to that emitted by the WT fusion protein, corresponding to higher deposition, was observed both for the Dutch and Artic mutations, whereas the Flemish fusion protein was more soluble than the WT form (Table 1). The results correspond closely with those documented in the literature, validating the approach used.

Discussion

The method used in this study is able to precisely connect the fluorescence emission of the GFP reporter to the aggregation propensity of the fused Aβ42 peptide. It has been shown that native GFP fused to aggregation-prone regions of yeast prions can be incorporated into aggregated amyloid structures and still fluoresce [41]. This is not the case in our study, where the reduced fluorescence emission observed for Aβ42–GFP variants with high aggregation propensities result from the inability of the GFP moiety to reach the native conformation from an initially unfolded state after its recombinant synthesis and before the aggregation event takes place. Coincidentally, an independent study has proven that the fluorescence of cells expressing C-terminal mutants of Aβ42 fused to GFP also correlates with protein aggregation [35].

In other protein models, the productive folding of the downstream GFP protein domain has been directly related to the folding performance of the upstream protein when over-expressed in *E. coli* [25]. However, aggregation of Aβ42 peptide is assumed to occur by direct self-assembly from an ensemble of unstructured conformations [42]. Hence, the observed changes in fluorescence emission should be related mainly to differences in the intrinsic aggregation properties of the different Aβ42 mutants, rather than to significant variations in their folding abilities. In this sense, the system used resembles the β-galactosidase complementation solubility assay which relies on intermolecular self-assembly rather than on folding properties of the protein fusions [43].

A direct conclusion from the data in Fig. 1 is that hydrophobic residues in the CHC of Aβ42 provide in general higher aggregation propensities than polar ones. The highly significant correlation observed between the residues' polarity and aggregation propensity confirms that an increased hydrophobicity usually leads to increased aggregation [32]. This is also evident from the observation that restoring the levels of hydrophobicity by mutation of the highly fluorescent mutant F19D (a double mutant F19D and E22F, see Supplementary material) results in a considerable decrease in fluorescence emission. Overall, the result is that the hydrophobicity in the CHC and adjacent positions of Aβ42 controls, at least partially, its deposition capabilities.

Aggregation, like protein folding, is thought to be determined by a balance of forces. Our analysis indicates that for Aβ42 CHC, and in addition to polarity, secondary structure propensities would modulate aggregation rates, as shown by the significant correlation found between β-sheet global tendency and aggregation in position 19. This supports the idea that the sequence tendency to promote aggregation is also related to its ability to form β-sheet strands from an unstructured conformation which may further favour the self-assembly into polymeric species by intermolecular bonding of the extended β-strands. Because hydrogen bonding within β-structure and hydrophobic interactions between side chains are likely to be the major stabilizing interactions within aggregates, increases in the propensities for such interactions are likely to enhance the rate at which aggregation occurs. Overall, the aggregation trends observed for the different side chains are in good agreement with those reported for protein models not related to disease [42] and, more importantly, with those described for both natural and synthetic/engineered Aβ42 mutants [44]. According to this, the additive combination of hydrophobicity, β-sheet global tendency and charge in the simple equation developed by Dobson and coworkers predicts with great accuracy the changes in fluorescence emission of mutants in position 19 of Aβ42. Our results suggest that, as Aβ42 is a mostly unstructured peptide, it is likely that simple physicochemical properties of the polypeptide chain might govern its aggregation propensity, lending support to the idea that common principles could underlie the aggregation of peptides and proteins, at least from unstructured states [42].

Traditionally, Pro has been the default substitution aimed at disrupting amyloid fibril formation, mainly because it disfavors local β-sheet folds, destabilizing the pathogen Aβ42 conformation [45]. It has been shown that the F19P mutation strongly reduces the incorporation of synthetic Aβ42 into amyloids [46]. Surprisingly, in our system the F19P mutant emits lower fluorescence than the F19D substitution, which we also show to block amyloid formation. This discrepancy may be understood considering that, although proline is very destabilizing for the fine β-sheet architecture of highly ordered and packed polypeptides in amyloid fibrils, it probably plays a more moderate role in less ordered aggregates, in which hydrophobicity appears to be the main driving force for aggregation. According to our analysis, the high reduction in aggregation propensity produced by the F19D mutation should be attributed to both a highly reduced hydrophobicity and β-sheet tendency in the CHC of the mutant protein. Interestingly enough, Street and Mayo have shown that Asp is the residue with the lowest theoretical and experimental β-sheet propensity (Gly and Pro could not be analysed) [47]. Charge would probably also influence the aggregation properties of F19D by increasing the net charge of the polypeptide. This observation may be biologically relevant, since chemical modifications of aspartate, such as isomerization, have been reported as examples of the very few post-translational modifications found in amyloid proteins isolated from amyloid deposits [48] and it has been shown that formation of isoaspartate increases the degree of fibril formation from Aβ protein *in vitro* [49]. Moreover, mutations of Asp residues result in increased amyloidogenicity in diseases caused by gelsolin, transthyretin, prion protein, lysozyme and immunoglobulin light chain (Bence–Jones) deposition

[50]. It has been shown recently that protein isoaspartate methyltransferase (PIMT) is a multicopy suppressor of protein aggregation in bacteria [51] and more interestingly that PIMT-deficient mice manifested neurodegenerative changes concomitant with the accumulation of L-isoaspartate in the brain [52]. In our study, both the data obtained *in vivo* using the Aβ42–GFP system and Eqn 1 advanced the strongly reduced amyloidogenic and cytotoxic abilities (see Supplementary material) of the F19D mutation, which were later confirmed by analysis of the 42-residue synthetic peptide. Taken together, the data suggest that Asp substitutions should be taken into account when new anti-aggregation strategies are designed.

The *in vivo* results obtained here in a prokaryotic background closely reproduce the properties of the natural Aβ occurring peptides bearing mutations related to early onset FAD. The increased fluorescence emitted by the Flemish mutant (A21G) is in complete agreement with the reduced rate of fibrillogenesis observed in humans, which may facilitate the diffusion or transport of the peptide from the brain parenchyma into the cerebral blood vessels, providing an explanation for the angiopathy and hemorrhagic components characteristic of Flemish disease. In contrast, the Dutch (E22Q) mutation results in a significant decrease in fluorescence emission respect to WT–GFP fusion, indicative of an increased aggregation propensity, which corresponds with its extensive aggregation ability in *in vitro* studies and the clinical evidence that Dutch patients are diagnosed as hereditary cerebral haemorrhage with amyloidosis. Finally, the recently described Arctic mutation (E22G) also results in a decrease in fluorescence emission in our analysis. This is coincident with the finding of increased protofibril formation and decreased Aβ plasma levels in the Arctic AD, which may reflect an alternative pathogenic mechanism involving a rapid Aβ protofibril formation which leads to an accelerated build-up of insoluble Aβ intra- and/or extracellularly. Overall, our data indicate that the properties of CHC and nearby residues in Aβ42 are important for stabilizing interactions involved in aggregation. Thus, this region emerges as a rational target for the development of assembly inhibitors of Aβ42. According to this, antibodies directed specifically against this peptide region strongly inhibit aggregation and toxicity of Aβ, decreasing brain Aβ burden in mouse models [53]. In addition, the results herein, together with the recent demonstration of amyloid-like properties of bacterial aggregates [27], prompts the use of prokaryotic models to explore the molecular determinants of protein aggregation by means of simple biological systems.

Experimental procedures

Site-direct mutagenesis

The vector expressing the Aβ42-GFP fusion was a generous gift of W. Kim, C. Wurth and M. Hecht (Princeton University, NJ, USA). Site-directed mutagenesis was performed using the QuickChange kit from Stratagene (La Jolla, CA, USA) according to the procedure recommended by the manufacturer. Forward and reverse primers were designed to change residues in positions 19, 21 and 22 of Aβ42. All constructs were verified by DNA sequencing. The WT and the mutated vectors were transformed into competent BL21 (DE3) cells. Cells were plated onto Luria–Bertani agar containing 50 $\mu g \cdot mL^{-1}$ kanamycin.

Expression of Aβ42–GFP mutants

BL21(DE3) cells harbouring WT or mutant Aβ42–GFP fusions were grow at 37 °C in Luria–Bertani medium containing 35 $\mu g \cdot mL^{-1}$ kanamycin. After 4 h, protein expression was induced with 1 mM isopropyl thio-β-D-galactoside. Cultures were grown for 3 h more, cells were then allowed to stand at 4 °C overnight to ensure fluorescence equilibrium and harvested by centrifugation. Expression of Aβ–GFP fusion proteins was monitored by SDS/PAGE using a 12% (w/v) gel.

Fluorescence measurements

Emission spectra of cells expressing WT and mutant Aβ42–GFP were measured on a Perkin Elmer 650-40 spectrofluorimeter (Boston, MA, USA). Bacterial cultures were grown, induced, and incubated overnight at 4 °C. Cells were diluted with 10 mM Tris/HCl pH 7.5 to an $A_{600} = 0.3$ and kept on ice until analysis. The fluorescence emission spectrum of the cell suspension was recorded from 500 to 600 nm, using an excitation wavelength of 450 nm (emission and excitation slits widths 5 mm). Data were corrected for buffer signals. At least three different scans were averaged for each protein sample.

Characterization of synthetic peptides

Wild-type and mutant Aβ42 synthetic peptides were obtained from American Peptide Company (Sunnyvale, CA, USA). Peptide samples were diluted in NH_4OH 0.02% to obtain a stock which was further diluted to the assay concentration in NaCl/P_i pH 7.5.

Circular dichroism spectra in the far UV region were obtained by using a UV-vis Jasco 715 spectro-polarimeter. Spectra were recorded at 25 °C at a peptide concentration ranging from 12.5 to 125 μM using a cell with a path length of 0.1 mm. Twenty scans were averaged to obtain each spectrum.

Peptides were tested for Congo red binding by spectroscopic band-shift assay as described by Klunk [54]. Peptides at 70 μM in NaCl/P$_i$ were incubated for 5 days at 25 °C. Aliquots of 50 μL peptide solutions were diluted in 950 μL of reaction solution (5 mM sodium phosphate/150 mM NaCl pH 7.0) containing 5 μM CR. Samples were equilibrated 5 min at 25 °C before analysis. Absorption spectra were collected together with that of a negative control of dye in absence of peptide on a CARY-100 Varian spectrophotometer (Les Ulis Cedex, France).

Thioflavin-T binding assays were carried out using aliquots of 20 μL drawn from 50 μM peptide samples in NaCl/P$_i$ incubated as indicated above. Aliquots were diluted into buffer (50 mM GlyNaOH pH 8.5) containing 100 μM Th-T, and adjusted to a final volume of 1 mL. Fluorescence emission spectra were recorded using an excitation wavelength fixed at 445 nm on a 650–40 Fluorescent Spectophotometer from Perkin-Elmer.

Aged peptide solutions were analysed by electron transmission microscopy. Samples were incubated at 37 °C during 48 h before measurements. Aliquots of 5–10 μL were placed on carbon-coated copper grids, and allowed to stand for 5 min. The grids were then washed and stained with 2% uranyl acetate for another 5 min prior to analysis using a HITACHI H-7000 transmission electron microscope operating at an accelerating voltage of 75 kV.

Calculation of changes in intrinsic polypeptidic properties

Δ Hydrophobicity is the change of hydrophobicity resulting from mutation and was calculated as previously described [42]. Briefly, Δ hydrophobicity = Hydr$_{wt}$–Hydr$_{mut}$ where Hydr$_{wt}$ and Hydr$_{mut}$ are the hydrophobicity values of the WT and mutant residues, respectively. The values of hydrophobicity for all 20 amino acids are from the Kyte–Doolittle hydrophobicity scale [55] or those based on the partition coefficients from water to octanol [42].

The difference in the free energy change for the transition random coil to β-sheet resulting from mutation ($\Delta\Delta G_{\beta\text{-coil}}$) and the predicted change of free energy for the transition α-helix to random coil resulting from mutation ($\Delta\Delta G_{\text{coil-}\alpha}$) were calculated mainly as described [42]. Briefly, $\Delta\Delta G_{\beta\text{-coil}}$ = 13.64(P_β^{wt}-P_β^{mut}), where P_β^{wt} and P_β^{mut} are the β-sheet propensities of the wild-type and mutant residue, respectively (the values of β-sheet propensity for all 20 amino acids were based on the scale of Minor and Kim, and 13,64 is the conversion constant from the normalized scale to units of kJmol^{-1}). $\Delta\Delta G_{\text{coil-}\alpha}$ = RTln(P_α^{wt}/P_α^{mut}), where P_α^{wt} and P_α^{mut} are the predicted α helical propensities (helix percentages) of the WT and mutated sequences at the site of mutation, respectively (calculated using the AGADIR algorithm at http://www.embl-heidelberg.de/Services/serrano/agadir/agadir-start.html).

Δβ-sheet propensity is the global change on the sequence propensity to form β-sheet upon mutation and was calculated as: Δβ-sheet propensity = $\Delta\Delta G_{\beta\text{-coil}}$ + $\Delta\Delta G_{\text{coil-}\alpha.}$

Correlation of fluorescence with polypeptide intrinsic properties

Fluorescence was plotted against the predicted aggregation rates of the different polypeptides calculated from Eqn 1 (developed by Chiti *et al.* [42]) This approach assumes that β-sheet propensity, hydrophobicity and charge are independent factors, which affect the aggregation of a protein, in an additive manner.

$$\ln(\nu_{\text{mut}}/\nu_{\text{wt}}) = \text{A}\Delta\text{Hydrophobicity} + \text{B}\Delta\beta\text{-sheet propensity} + \text{C}\Delta\text{Charge} \quad (1)$$

where ν_{mut} and ν_{wt} correspond to the predicted aggregation rates of the mutant and WT sequences, respectively, and ΔCharge is the difference in the net charge of the polypeptide introduced by the mutation. A, B and C-values are constants determined experimentally from the analysis of a large set of mutants of Acylphosphatase [42].

Acknowledgements

This work has been supported by Grants BIO2001-2046 and BIO2004-05879 (Ministerio de Ciencia y Tecnología, MCYT, Spain), by the Centre de Referència en Biotecnologia (Generalitat de Catalunya, Spain), by PNL2004-40 (Universitat Autònoma de Barcelona (UAB)) and 2005SGR-00037 (AGAUR). SV is recipient of a 'Ramón y Cajal' contract awarded by the MCYT-Spain and cofinanced by the UAB.

References

1 Smith A (2003) Protein misfolding. *Nature* **426**, 883.

2 Tan SY & Pepys MB (1994) Amyloidosis. *Histopathology* **25**, 403–414.

3 Cohen FE & Kelly JW (2003) Therapeutic approaches to protein-misfolding diseases. *Nature* **426**, 905–909.

4 Rochet JC & Lansbury PT Jr (2000) Amyloid fibrillogenesis: themes & variations. *Curr Opin Struct Biol* **10**, 60–68.

5 Dobson CM (2003) Protein folding and misfolding. *Nature* **426**, 884–890.

6 Mayeux R (2003) Epidemiology of neurodegeneration. *Annu Rev Neurosc* **26**, 81–104.

7 Selkoe DJ (2001) Alzheimer's disease: genes, proteins & therapy. *Physiol Rev* **81**, 741–766.

8 Glenner GG & Wong CW (1984) Alzheimer's disease: initial report of the purification and characterization of a novel cerebrovascular amyloid protein. *Biochem Biophys Res Commun* **120**, 885–890.

9 Jarrett JT, Berger EP & Lansbury PT Jr (1993) The carboxy terminus of the beta amyloid protein is critical for the seeding of amyloid formation: implications for the pathogenesis of Alzheimer's disease. *Biochemistry* **32**, 4693–4697.

10 Nagele RG, Wegiel J, Venkataraman V, Imaki H, Wang KC & Wegiel J (2004) Contribution of glial cells to the development of amyloid plaques in Alzheimer's disease. *Neurobiol Aging* **25**, 663–674.

11 Dominguez DI & De Strooper B (2002) Novel therapeutic strategies provide the real test for the amyloid hypothesis of Alzheimer's disease. *Trends Pharmacol Sci* **23**, 324–330.

12 Suzuki N, Cheung TT, Cai XD, Odaka A, Otvos L Jr, Eckman C, Golde TE & Younkin SG (1994) An increased percentage of long amyloid beta protein secreted by familial amyloid beta protein precursor (beta APP717) mutants. *Science* **264**, 1336–1340.

13 Borchelt DR, Thinakaran G, Eckman CB, Lee MK, Davenport F, Ratovitsky T, Prada CM, Kim G, Seekins S, Yager D *et al.* (1996) Familial Alzheimer's disease-linked presenilin 1 variants elevate Abeta1–42/1–40 ratio in vitro & in vivo. *Neuron* **17**, 1005–1013.

14 Lemere CA, Lopera F, Kosik KS, Lendon CL, Ossa J, Saido TC, Yamaguchi H, Ruiz A, Martinez A, Madrigal L *et al.* (1996) The E280A presenilin 1 Alzheimer's mutation produces increased A beta 42 deposition & severe cerebellar pathology. *Nat Med* **2**, 1146–1150.

15 Mann DM, Iwatsubo T, Nochlin D, Sumi SM, Levy-Lahad E & Bird TD (1997) Amyloid (Abeta) deposition in chromosome 1-linked Alzheimer's disease: the Volga German families. *Ann Neurol* **41**, 52–57.

16 Lott IT & Head E (2005) Alzheimer's disease & Down syndrome: factors in pathogenesis. *Neurobiol Aging* **26**, 383–389.

17 Hilbich C, Kisters-Woike B, Reed J, Masters CL & Beyreuther K (1992) Substitutions of hydrophobic amino acids reduce the amyloidogenicity of Alzheimer's disease beta A4 peptides. *J Mol Biol* **228**, 460–473.

18 Wood SJ, Wetzel R, Martin JD & Hurle MR (1995) Prolines & amyloidogenicity in fragments of the Alzheimer's peptide beta/A4. *Biochemistry* **34**, 724–730.

19 Balbach JJ, Ishii Y, Antzutkin ON, Leapman RD, Rizzo NW, Dyda F, Reed J & Tycko R (2000) Stabilities and conformations of Alzheimer's beta-amyloid peptide oligomers (Abeta 16–22, Abeta 16–35 & Abeta 10–35): Sequence effects. *Biochemistry* **39**, 13748–13759.

20 Findeis MA, Musso GM, Arico-Muendel CC, Benjamin HW, Hundal AM, Lee JJ, Chin J, Kelley M, Wakefield J, Hayward NJ *et al.* (1999) Modified-peptide inhibitors of amyloid beta-peptide polymerization. *Biochemistry* **38**, 6791–6800.

21 Williams AD, Portelius E, Kheterpal I, Guo JT, Cook KD, Xu Y & Wetzel R (2004) Mapping Abeta amyloid fibril secondary structure using scanning proline mutagenesis. *J Mol Biol* **335**, 833–842.

22 Morimoto A, Irie K, Murakami K, Masuda Y, Ohigashi H, Nagao M, Fukuda H, Shimizu T & Shirasawa T (2004) Analysis of the secondary structure of beta-amyloid (Abeta42) fibrils by systematic proline replacement. *J Biol Chem* **279**, 52781–52788.

23 Bitan G, Vollers SS & Teplow DB (2003) Elucidation of primary structure elements controlling early amyloid beta-protein oligomerization. *J Biol Chem* **278**, 34882–34889.

24 Teplow DB, Lomakin A, Benedek GB, Kirschner DA & Walsh DM (1997) Effects of Beta-protein mutations on amyloid fibril nucleation & elongation. In *Alzheimer's Disease: Biology, Diagnosis and Therapeutics* (K Iqbal, B Winblad, T Nishimura, M Takeda & HM, eds), p. 311. John Wiley and Sons Ltd, Chichester, UK.

25 Waldo GS, Standish BM, Berendzen J & Terwilliger TC (1999) Rapid protein-folding assay using green fluorescent protein. *Nature Biotechnol* **17**, 691–695.

26 Wurth C, Guimard NK & Hecht MH (2002) Mutations that reduce aggregation of the Alzheimer's Abeta42 peptide: an unbiased search for the sequence determinants of Abeta amyloidogenesis. *J Mol Biol* **319**, 1279–1290.

27 Carrio M, Gonzalez-Montalban N, Vera A, Villaverde A & Ventura S (2005) Amyloid-like properties of bacterial inclusion bodies. *J Mol Biol* **347**, 1025–1037.

28 Fasman GD, Perczel A & Moore CD (1995) Solubilization of beta-amyloid-(1–42)-peptide: reversing the beta-sheet conformation induced by aluminum with silicates. *Proc Natl Acad Sci USA* **92**, 369–371.

29 LeVine H (1993) Thioflavine T interaction with synthetic Alzheimer's disease beta-amyloid peptides: detection of amyloid aggregation in solution. *Protein Sci* **2**, 404ñ410.

30 Bucciantini M, Giannoni E, Chiti F, Baroni F, Formigli L, Zurdo J, Taddei N, Ramponi G, Dobson CM & Stefani M (2002) Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature* **416**, 507–511.

31 Ruth L, Eisenberg D & Neufeld EF (2000) alpha-L-iduronidase forms semi-crystalline spherulites with amyloid-like properties. *Acta Cryst D* **56**, 524–528.

32 Fink AL (1998) Protein aggregation: folding aggregates, inclusion bodies and amyloid. *Fold Des* **3**, R9–R23.

33 Chiti F, Calamai M, Taddei N, Stefani M, Ramponi G & Dobson CM (2002) Studies of the aggregation of mutant proteins in vitro provide insights into the genetics of amyloid diseases. *Proc Natl Acad Sci USA* **99**, 16419–16426.

34 Lopez de la Paz M, Goldie K, Zurdo J, Lacroix E, Dobson CM, Hoenger A & Serrano L (2002) De novo designed peptide-based amyloid fibrils. *Proc Natl Acad Sci USA* **99**, 16052–16057.

35 Kim W & Hecht MH (2005) Sequence determinants of enhanced amyloidogenicity of Alzheimer's Abeta42 peptide relative to Abeta40. *J Biol Chem* **280**, 35069–35076.

36 Hendriks L, van Duijn CM, Cras P, Cruts M, Van Hul W, van Harskamp F, Warren A, McInnis MG, Antonarakis SE, Martin JJ *et al.* (1992) Presenile dementia and cerebral haemorrhage linked to a mutation at codon 692 of the beta-amyloid precursor protein gene. *Nat Genet* **1**, 218–221.

37 Levy E, Carman MD, Fernandez-Madrid IJ, Power MD, Lieberburg I, van Duinen SG, Bots GTAM, Luyendijk W & Frangione B (1990) Mutation of the Alzheimer's disease amyloid gene in hereditary cerebral hemorrhage, Dutch type. *Science* **248**, 1124–1126.

38 Kamino K, Orr HT, Payami H, Wijsman EM, Alonso E, Pulst SM, Anderson L, O'dahl S, Nemens E, White JA *et al.* (1992) Linkage and mutational analysis of familial Alzheimer's disease kindreds for the APP gene region. *Am J Hum Genet* **51**, 998–1014.

39 van Nostrand WE, Melchor JP, Cho HS, Greenberg SM & Rebeck GW (2001) Pathogenic effects of D23N Iowa mutant amyloid beta-protein. *J Biol Chem* **276**, 32860–32866.

40 Nilsberth C, Westlind-Danielsson A, Eckman CB, Condron MM, Axelman K, Forsell C, Stenh C, Luthman J, Teplow DB, Younkin SG *et al.* (2001) The 'Arctic' APP mutation (E693G) causes Alzheimer's disease by enhanced Abeta protofibril formation. *Nat Neurosci* **4**, 887–893.

41 Baxa U, Speransky V, Steven AC & Wickner RB (2002) Mechanism of inactivation on prion conversion of the Saccharomyces cerevisiae Ure2 protein. *Proc Natl Acad Sci USA* **99**, 5253–5260.

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

43 Wigley WC, Stidham RD, Smith NM, Hunt JF & Thomas PJ (2001) Protein solubility and folding monitored in vivo by structural complementation of a genetic marker protein. *Nat Biotechnol* **19**, 131–136.

44 Sanchez de Groot N, Pallares I, Aviles FX, Vendrell J & Ventura S (2005) Prediction of 'hot spots' of aggregation in disease-linked polypeptides. *B M C Struct Biol* **5**, 18.

45 Esler WP, Stimson ER, Ghilardi JR, Lu YA, Felix AM, Vinters HV, Mantyh PW, Lee JP & Maggio JE (1996) Point substitution in the central hydrophobic cluster of a human beta-amyloid congener disrupts peptide folding and abolishes plaque competence. *Biochemistry* **35**, 13914–13921.

46 Bernstein SL, Wyttenbach T, Baumketner A, Shea JE, Bitan G, Teplow DB & Bowers MT (2005) Amyloid beta-protein: monomer structure and early aggregation states of Abeta42 and its Pro19 alloform. *J Am Chem Soc* **127**, 2075–2084.

47 Street AG & Mayo SL (1999) Intrinsic beta-sheet propensities result from van der Waals interactions between side chains and the local backbone. *Proc Natl Acad Sci USA* **96**, 9074–9076.

48 Shapira R, Austin GE & Mirra SS (1988) Neuritic plaque amyloid in Alzheimer's disease is highly racemized. *J Neurochem* **50**, 69–74.

49 Velazquez P, Cribbs DH, Poulos TL & Tenner AJ (1997) Aspartate residue 7 in amyloid beta-protein is critical for classical complement pathway activation: implications for Alzheimer's disease pathogenesis. *Nature Med* **3**, 77–79.

50 Benson MD (1995) Amyloidosis. In *Metabolic Basis of Inherted Disease* (Scriver, C R, Beandet, A L, Sly, WS & Valle, D, eds), pp. 4159. McGraw-Hill Inc, New York.

51 Kern R, Malki A, Abdallah J, Liebart JC, Dubucs C, YuMH & Richarme G (2005) Protein isoaspartate methyltransferase is a multicopy suppressor of protein aggregation in Escherichia coli. *J Bacteriol* **187**, 1377–1383.

52 Shimizu T, Matsuoka Y & Shirasawa T (2005) Biological significance of isoaspartate and its repair system. *Biol Pharm Bull* **28**, 1590–1596.

53 Dodart JC, Bales KR, Gannon KS, Greene SJ, DeMattos RB, Mathis C, DeLong CA, Wu S, Wu X, Holtzman DM & Paul SM (2002) Immunization reverses memory deficits without reducing brain Abeta burden in Alzheimer's disease model. *Nat Neurosci* **5**, 452–457.

54 Klunk WE, Pettegrew JW & Abraham DJ (1989) Quantitative evaluation of congo red binding to amyloid-like proteins with a beta-pleated sheet conformation. *J Histochem Cytochem* **37**, 1273–1281.

55 Kyte J & Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**, 105–132.

Supplementary material

The following supplementary material is available online:

Mutagenesis of the central hydrophobic cluster in Aβ42 Alzheimer's peptide.

Fig. S1. SDS/PAGE analysis of the expression of wild-type and selected mutant Aβ42.

Fig. S2. Experimental fluorescence of WT, D19F and the double mutants F19D and E22F.

Fig. S3. Cytotoxicity of Aβ42 synthetic peptides.

This material is available as part of the online article from http://www.blackwell-synergy.com

Ile-Phe Dipeptide Self-Assembly: Clues to Amyloid Formation

Natalia Sánchez de Groot,* Teodor Parella,† Francesc X. Aviles,*‡ Josep Vendrell,*‡ and Salvador Ventura*‡
*Departament de Bioquímica i Biologia Molecular, †Servei de Ressonància Magnètica Nuclear, and ‡Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain

ABSTRACT Peptidic self-assembled nanostructures are said to have a wide range of applications in nanotechnology, yet the mechanistic details of hierarchical self-assembly are still poorly understood. The Phe-Phe recognition motif of the Alzheimer's Aβ peptide is the smallest peptide able to assemble into higher-order structures. Here, we show that the Ile-Phe dipeptide analog is also able to self-associate in aqueous solution as a transparent, thermoreversible gel formed by a network of fibrillar nanostructures that exhibit strong birefringence upon Congo red binding. Besides, a second dipeptide Val-Phe, differing only in a methyl group from the former, is unable to self-assemble. The detailed analysis of the differential polymeric behavior of these closely related molecules provides insight into the forces triggering the first steps in self-assembly processes such as amyloid formation.

INTRODUCTION

Successful synthesis of organized supramolecular assemblies is a fundamental step toward the release of new materials or functional supramolecular devices. The controlled self-assembly of biomolecular structures, preferably from the simplest building blocks possible, is therefore of great interest (1–3). Gels represent new soft biocompatible materials that have numerous potential applications in fields like biomaterials, biosensors, tissue engineering, and drug delivery (4–6). Under appropriate conditions, self-assembled arrays of natural and designed proteins and peptides are often observed to trap bulk solvent and result in the formation of transparent gels (7–9). Peptides have emerged as promising gelling compounds since their self-assembly results from the interplay of several weak interactions, such as hydrogen bonding, electrostatics, and hydrophobic forces, which finally organize the monomeric components and lead to the generation of long, noncovalent, supramolecular assemblies (10). These noncovalent forces are reminiscent of those driving amyloid fibril formation, which result, both in vivo and in vitro, in the formation of highly ordered supramolecular assemblies from initially monomeric species (11,12). Fragments of the major proteins involved in Alzheimer's disease, i.e., Tau and Aβ42 peptide, have been shown to act as gelators in vitro, whereas microscopic characterization of the gels has revealed the presence of fibril networks (13,14).

We and other authors have recently shown that specific short stretches in proteins are responsible for their aggregating behavior (15–17) and, in agreement with this observation, several short peptides of amyloidogenic proteins have been shown to form supramolecular structures, indistinguishable from those formed by the complete polypeptide chains (18,19).

Besides their easy design and synthesis, short peptides are both excellent model systems for the study of biological self-assembly and ideal building blocks for the production of a wide range of biological materials. The dipeptide NH_2-Phe-Phe-COOH, described as the smallest peptide able to assemble into higher-order structures (20), corresponds to residues 19 and 20 of the central hydrophobic cluster (CHC) of the highly amyloidogenic peptide Aβ42. Position 19 has been shown to strongly affect the assembly and aggregation of Aβ (21). In a recent work, we substituted Phe^{19} with the other 19 proteinogenic amino acids and assayed the effect of these single mutations on Aβ42's aggregation (22). All substitutions, with the exception of Phe^{19}Ile, resulted in peptides with decreased aggregation propensities relative to that of the wild-type molecule. Thus, an interest arose to determine the molecular properties of the dipeptide NH_2-Ile-Phe-COOH (Fig. 1), an analog of the diphenylalanine element shown to self-assemble in vitro. Here, we show that the Ile-Phe dipeptide self-associates to form a transparent, thermoreversible gel formed by a network of fibrillar nanostructures in water. Besides, a second dipeptide NH_2-Val-Phe-COOH, differing only in a methyl group from the former, is unable to self-assemble, in agreement with the lower aggregation propensity reported for Phe^{19}Val Aβ42 relative to that of the Phe^{19}Ile version (22). The detailed analysis of the differential self-association capability of these two molecules provides clues for the understanding of hierarchical self-assembly.

METHODS

General methods and materials

Lyophilized NH_2-Ile-Phe-COOH dipeptide and NH_2-Val-Phe-COOH dipeptide with >99% purity were obtained from Bachem (Bubendorf, Switzerland). Dipeptide samples were diluted in 1,1,1,3,3,3 hexafluoro-2-propanol to obtain stocks of 100 mg/ml and 200 mg/ml, which were further diluted to the assay concentration in H_2O, except for NMR and Fourier transform infrared (FTIR) assays, where the samples were diluted in deuterium oxide

Submitted September 6, 2006, and accepted for publication November 16, 2006.

Address reprint requests to Salvador Ventura, Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain. E-mail: salvador.ventura@uab.es.

0006-3495/07/03/1732/10 $2.00

doi: 10.1529/biophysj.106.096677

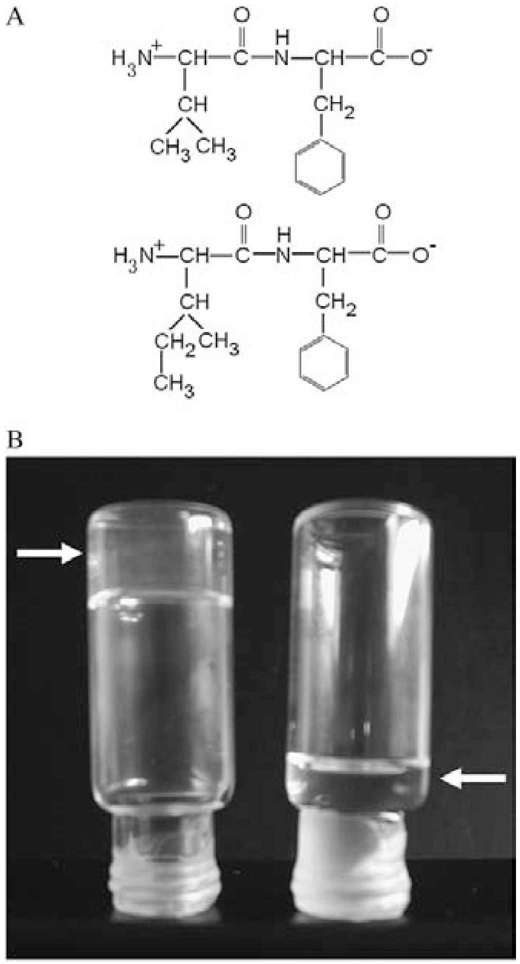


FIGURE 1 Gelation of Ile-Phe dipeptide. (*A*) Structures of the dipeptides compared in this work. NH_3^+-Val-Phe-COO^- (*upper*) and NH_3^+-Ile-Phe-COO^- (*lower*). (*B*) Photograph of 2% (w/v) dipeptide samples under blue light. NH_3^+-Ile-Phe-COO^- forms a gel (*left*) whereas NH_3^+-Val-Phe-COO^- remains in solution (*right*).

(D_2O). 1,1,1,3,3,3 hexafluoro-2-propanol and 2-(*p*-toluidinylnaphthalene)-6-sulfonate (TNS) were purchased from Sigma (St. Louis, MO). D_2O enriched >99.97% in isotope D_2O ($d = 1.11$) was purchased from SDS (13124, Peypin, France).

Light absorbance at 360 nm

The turbidity of the different dipeptide samples at each temperature was measured monitoring the absorbance at 360 nm on a CARY-400 Varian spectrophotometer (Les Ulis, France). To study the dependence of peptide self-assembly on concentration, the turbidity was measured at 293 K. To study the dependence of peptide assembly state on the temperature, each sample was first heated to 333 K or cooled to 283 K before measuring the reassembly or disassembly, respectively. These samples were subsequently cooled or heated in 5-K stages and equilibrated for 15 min before measuring the turbidity at each assayed temperature.

NMR

NMR experiments were collected in a 500-MHz Avance Bruker spectrometer (Berlin, Germany) equipped with a triple-resonance TXI probehead. High-resolution ^{1}H NMR spectra were recorded for several Ile-Phe and Val-Phe dipeptide samples at different concentrations and different temperatures. The samples used consisted of 0.05%, 0.1%, 0.5%, 1%, and 2% (w/v) of dipeptide dissolved in D_2O from stock solutions. Spectra were also collected in the range 295–330 K to observe the temperature dependence of each individual sample.

Microscopy

Dipeptide samples (1.5%, w/v) were placed on carbon-coated copper grids and left for 5 min. The grids were then stained with 2% (w/v) uranyl acetate for another 2 min before analysis using a Hitachi (Tokyo, Japan) H-7000 transmission electron microscope operating at an accelerating voltage of 75 kV. A sample of 2% (w/v) Ile-Phe gel smeared on a 1-cm slide was allowed to dry at room temperature, followed by gold coating, before being imaged on a Hitachi S-570 scanning electron microscope.

Congo red binding

Congo red (CR) was diluted in a buffer containing 5 mM sodium phosphate and 150 mM NaCl, pH 7.0, to obtain a stock of 100 μM CR. A 2% (w/v) Ile-Phe gel was formed in the presence of 5 μM CR final concentration. A gel sample was placed on a microscope slide and sealed. The CR birefringence was detected under cross-polarized light using an optic microscope (Leica DMRB, Heidelberg, Germany).

Absorption and fluorescent spectra of Phe

The Phe fluorescence emission spectra of the dipeptide samples were recorded in a Perkin–Elmer (Wellesley, MA) 650–40 fluorescence spectrophotometer. The samples were excited at 250 nm and the emission between 260 nm and 400 nm was measured. Both excitation and emission slits were set at 10 nm. The absorption spectrum of Phe was measured between 230 nm and 330 nm on a CARY-400 Varian spectrophotometer.

FTIR

Diluted, gelled, and air-dried dipeptide samples were used for FTIR spectroscopy analysis. Exchangeable hydrogen atoms were replaced by deuterium by dissolving the dipeptide stocks in D_2O. Infrared spectra were recorded with an FTS-6000 FT-IR spectrophotometer (BioRad, Hemel Hempstead, UK) equipped with a liquid nitrogen-cooled mercury/cadmium telluride detector and purged with a continuous flow of nitrogen gas or with a Bruker Tensor 27 FT-IR spectrometer. For each spectrum, 200 interferograms were collected and averaged. In every case, the buffer spectrum was subtracted and the baseline corrected. Second derivatives of the spectra were used to determine the frequencies at which the different spectral components were located. To monitor the effect of pH on self-assembly, we measured the pH established by the dipeptides themselves upon dilution in H_2O, pH 5.8. This pH was either increased to pH 12.0 by addition of 1 N NaOH or lowered to pH 2.0 by addition of 1 N HCl to test the effect of N- and C-terminal group ionization in Ile-Phe self-assembly.

TNS binding

The fluorescence emission spectra of TNS with the dipeptide samples were recorded at 293 K in a Perkin-Elmer 650-40 fluorescence spectrophotometer. TNS was diluted in H_2O to obtain a 1 mM stock solution. The samples were excited at 323 nm and the fluorescence emission was measured between 350 nm and 550 nm. Both excitation and emission slits were set at 10 nm. To follow the Ile-Phe dipeptide kinetic self-assembly by TNS binding, a 2% Ile-Phe sample in a solution containing a final 10 μM TNS concentration was heated to 333 K and then cooled gradually to 278 K. The sample was excited at 323 nm and the fluorescence emission at 423 nm was monitored.

Light scattering

Light scattering of a 2% Ile-Phe sample was measured using a Perkin-Elmer 650-40 fluorescence spectrophotometer. The sample was excited at 360 nm and the emission at the same wavelength was monitored. To follow the Ile-Phe dipeptide kinetic self-assembly, the sample was heated and cooled as in the TNS binding assay.

RESULTS

General observations

Lyophilized dipeptides could be dissolved at very high concentrations (200 mg/ml) in 1,1,1,3,3,3 hexafluoro-2-propanol. Although both peptides appeared to be highly soluble in the organic solvent, a rapid assembly into macroscopic structures was observed visually for the Ile-Phe peptide soon after dilution on H_20 at final concentrations $>0.5\%$, w/v. At 1%, w/v, the solution begins to gelate, forming a solid gel above 1.5% (w/v) final concentration (Fig. 1). Surprisingly, the Val-Phe solution remained liquid in all assayed conditions (Fig. 1).

Dependence of dipeptide self-assembly on the concentration

Gelation usually represents a macroscopic manifestation of a molecular self-assembly process (10), thus suggesting the formation of high aspect ratio nanostructures by the Ile-Phe dipeptide. To better quantify the dependence of peptide self-assembly on the concentration, we monitored the changes in light absorbance at 360 nm (Fig. 2 *A*) and recorded ^{1}H-NMR spectra of solutions of both peptides at different concentrations (Fig. 2 *B*). The absorbance of Ile-Phe solutions at 360 nm is highly dependent on the peptide concentration, producing a sigmoid curve in which the transition between the soluble and polymerized states occurs at 1.1% (w/v) peptide concentration. No increase in absorbance was detected for the Val-Phe solutions even at high peptide concentrations.

As expected, well-resolved sharp peaks can be clearly seen in the ^{1}H-NMR spectra of both dipeptides at 0.02%, w/v, in aqueous solution (Fig. 2 *B*), indicating their monomeric status. However, as peptide concentration increases, the signals in the spectrum of the Ile-Phe solutions progressively broaden and most NMR peaks become unresolved, indicating a decrease in molecular motion due to the formation of the supramolecular structure. In good agreement with the visible (VIS) spectroscopic data, this effect is especially observable at peptide concentrations $>0.5\%$, w/v. The spectra of Val-Phe solutions display well-resolved signals at all concentrations assayed, confirming that the dipeptide is unable to self-associate into higher-order structures in water (Fig. 2 *B*).

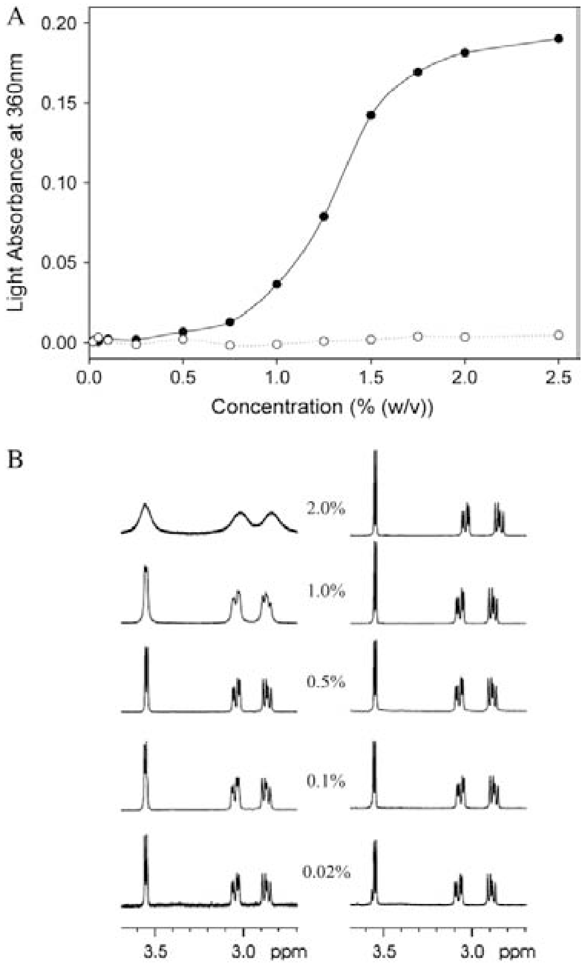


FIGURE 2 Dependence of peptide self-assembly on the concentration. (*A*) Turbidity of NH_3^+-Ile-Phe-COO^- (*solid circles*) and NH_3^+-Val-Phe-COO^- (*open circles*) at different peptide concentrations. (*B*) ^{1}H NMR spectra of NH_3^+-Ile-Phe-COO^- (*left*) and NH_3^+-Val-Phe-COO^- (*right*) at different peptide concentrations (w/v).

Structure of the Ile-Phe gel

The nanometric structures formed by the Ile-Phe dipeptide correspond to well-ordered, fibrillar, and elongated assemblies as seen by transmission electron microscopy (TEM) analysis with negative staining (Fig. 3), with almost no presence of amorphous aggregates. This is in contrast with other peptide assemblies, such as amyloid fibrils, in which molecules are easily trapped in kinetically stable arrangements of different topology, usually resulting in a mixture of structured and nonordered material (23). The formed structures are highly ordered and homogeneous, without branching. This can be also observed using scanning electron microscopy (SEM) (Fig. 3). The fibrils display a consistent width of ~55 nm, which is clearly larger than that reported for typical amyloid fibrils but similar in diameter to the amyloid-like self-assembled peptide nanotubular structures described for diphenylalanine (20). In the TEM images the fibrillar structures appear to be quite transparent to the beam

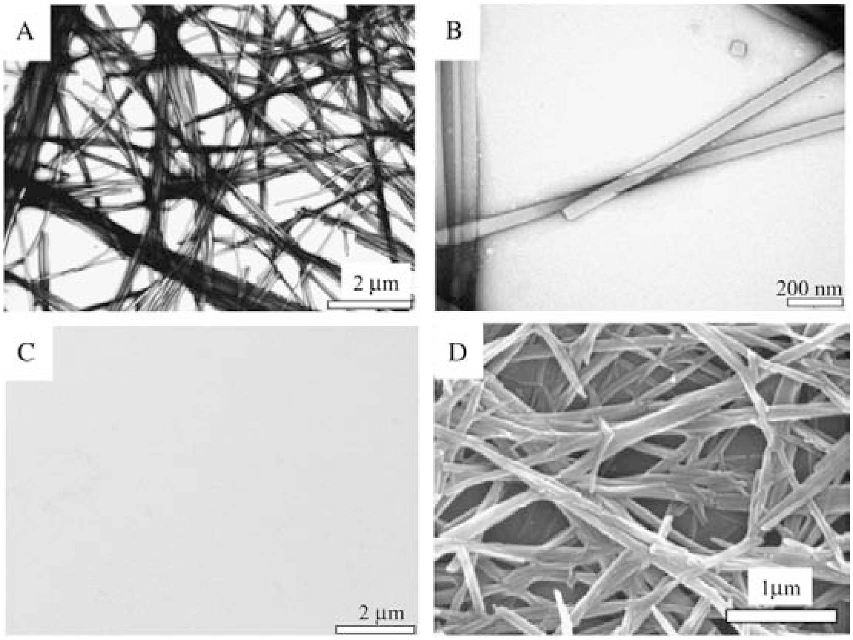


FIGURE 3 Electron microscopy (EM) images of dipeptide samples. (*A* and *B*) Transmission EM images of a 1.5% (w/v) NH_3^+-Ile-Phe-COO$^-$ sample at increasing magnification. (*C*) Transmision EM image of a 1.5% (w/v) NH_3^+-Val-Phe-COO$^-$ sample. (*D*) Scanning EM image of a gel formed by 2% (w/v) NH_3^+-Ile-Phe-COO$^-$.

of electrons (Fig. 3), which could suggest that they are more or less hollow. The fibrils are very long (several micrometers) and usually appear to be laterally associated. Only a small number of them appear to be slightly twisted, with most remaining linear, suggesting that they do not tend to adopt a twisted helical structure. A similar observation was made with the diphenylalanine nanotubular structure described by Reches and Gazit (20). The pack of fibrils entangles into a supramolecular network, which is expressed macroscopically as the observed gel. No ordered or amorphous aggregated material could be observed in solutions of the Val-Phe dipeptide by TEM analysis (Fig. 3).

To further determine the properties of the structures observed by TEM and SEM, Ile-Phe samples (2%, w/v) were stained with the amyloidophilic CR. The structures formed by the dipeptide show a strong green-gold birefringence upon incubation when illuminated under cross-polarized light (Fig. 4), indicating that the Ile-Phe dipeptide already contains all the molecular information needed to self-associate into regular structures that may be somehow similar to those in amyloid fibrils.

Dependence of Ile-Phe self-assembly on the temperature

Among other reasons, weak forces are useful for the construction of self-assembled materials because they allow reversibility. This property allows materials to respond to their environment by assembling and disassembling, an important factor in the design of ''smart materials'', sensors, or controlled-release devices (24,25). The so-called thermoresponsive materials are an especially interesting kind of nanostructure in which the association state of the building blocks depends on the temperature. To test the temperature dependence of the formation of supramolecular assemblies by the Ile-Phe dipeptide, we monitored the changes in light absorbance at 360 nm and recorded the ^{1}H-NMR spectra at different temperatures (Fig. 5). The nanostructures formed by the dipeptide at 2% (w/v) are sensitive to temperature changes, as indicated by the progressive decrease in absorbance observed when the temperature increases (Fig. 5 *A*). The melting curve is cooperative and fully reversible, indicating that the assembling and disassembling processes are occurring in a coordinated way, as expected for a self-associated structure in which the noncovalent interactions linking the blocks are progressively gained or lost. It may be observed that the dipeptide sample is solid at 293 K, becomes completely fluid above 313 K, and recovers its initial state upon cooling. The transition temperature depends on the concentration of the dipeptide in the sample, being 304 K for a 2%, w/v, and 299 K for a 1.5%, w/v, sample. The reversible assembly and disassembly of the Ile-Phe in response to changes in temperature was also confirmed by ^{1}H-NMR (Fig. 5 *B*). The low-temperature NMR signals become progressively better resolved as temperature increases, indicating higher mobility of the building blocks and, thus, disorganization of the supramolecular structures. The signal is broad again upon cooling and the spectrum becomes indistinguishable from the initial one, indicating reassembly of the fibrillar organization. No dependence on the temperature was detected for a 2%, w/v, Val-Phe sample either by VIS-spectroscopy or ^{1}H-NMR spectroscopy.

Molecular interactions implied in the assembly process

FTIR spectra were obtained to identify intermolecular interactions in the Ile-Phe gel and solution states (Table 1). At

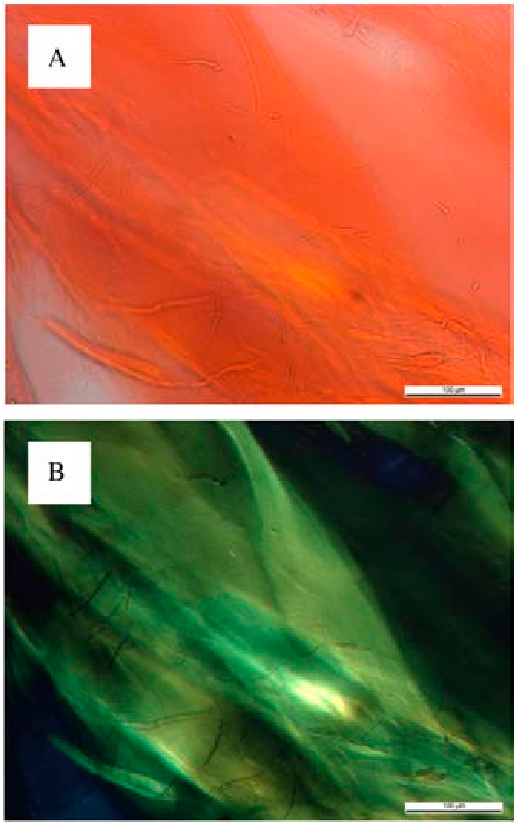


FIGURE 4 Congo red staining and birefringence of 2% (w/v) NH_3^+-Ile-Phe-COO^-. (*A*) Dipeptide stained with Congo red and observed at 40× magnification. (*B*) Same sample observed between crossed polarizers, displaying the green birefringence characteristic of amyloid structures.

0.1% (w/v) peptide concentration, an NH band at 3398 cm^{-1} and an amide I band at 1662 cm^{-1} corresponding to non-hydrogen-bonded NH and CO functionalities were detected. The absence of a CO stretching band above 1700 cm^{-1}, together with the additional detection of a strong vibrational signal at 1598 cm^{-1} resulting from the asymmetric stretching of the C-terminal COO^- group, indicates that all COO^- groups are deprotonated in the dipeptide solution, pH 5.8 (26–28). Surprisingly, the NH and amide I bands are still detected at positions corresponding to non-hydrogen-bonded groups in the polymerized state (2%, w/v), indicating that the amide bond is not involved in self-assembly in aqueous solution. By contrast, the signal at 1598 cm^{-1} suffers a strong downshift to 1570 cm^{-1}, a change in the position of the C-terminal COO^- group that has been previously observed in the vicinity of NH_3^+ groups (29). Since the dipeptide contains only a single rigid amide bond, the coupling between COO^- and NH_3^+ groups suggests a supramolecular structure stabilized by head-to-tail-interactions between dipeptides. According to the above-described observation of a gradual network disintegration into soluble and probably monomeric species upon heating, the COO^- band shifts again to higher wavenumbers (1597 cm^{-1}) when the temperature is increased to 323 K, indicating a disruption of the intermolecular interactions formed between the dipeptide molecules. The process is fully reversible and the spectrum recovers its original shape upon cooling. The COO^- vibrational downshift is not observed at intermediate peptide concentrations (0.8%, w/v) at which the presence of self-assembled species is already detected by VIS spectroscopy and ^{1}H-NMR, suggesting that the changes in the local environments of COO^- groups occur during or after the consolidation of the nanostructures. As expected for a non-self-assembling species, a COO^- band at 1598 cm^{-1} appears in a 2% (w/v) solution of the Val-Phe dipeptide. Lowering the pH of a gel solution to 2.0 by adding HCl or increasing it to 12.0 by adding NaOH further demonstrated the role played by the COO^- and NH_3^+ groups in maintaining a highly ordered nanostructure. In these conditions, the carboxyl and amino groups become, respectively, protonated and deprotonated and this results in the disintegration of the network and the formation of numerous amorphous aggregates after 1 –h of incubation (data not shown).

The typical amide I peak at ~1620–1640 cm,$^{-1}$ usually associated with the presence of β-sheet structures, was not detected in any of the samples analyzed. It has been shown that the loss of bound water in the gel formed by a Tau peptide results in increased formation of β-sheet structure in the gel and, finally, fibrillation (13). The FTIR spectra of a dehydrated gel sample showed no significant shift in the COO^- band. In contrast, a strong band at 3305 cm^{-1} in the NH region corresponding to hydrogen-bonded NHs could be detected, whereas the band corresponding to free NHs was minor. In the amide-I region the presence of additional bands at 1618 cm^{-1} and 1678 cm^{-1} is indicative of the formation of new β-sheet-like intermolecular bonds in the supramolecular structure upon loss of water.

Role of hydrophobicity in the assembly process

The previous data suggest that head-to-tail interactions between the amino and carboxyl-terminus of dipeptides stabilize the assembly of fibrils. However, the establishment of such intermolecular contacts could not be the initial driving force for Ile-Phe polymerization, as the Val-Phe dipeptide possesses exactly the same groups and remains in solution, most likely in the monomeric state, at high concentrations. For the same reason, the stacking between aromatic rings can be discarded as the main interaction promoting self-assembly in this particular peptide system. This is confirmed by several observations. First, the signals of aromatic protons in ^{1}H-NMR spectra of Ile-Phe are only slightly shifted upfield as the temperature is increased, whereas aromatic signals should be significantly shifted downfield after disassembly of intermolecular aromatic stacking interactions in an assembled peptide. Second, the absorption and fluorescent spectra of Phe in dilute samples (0.02%, w/v) and at near-transition concentration (1%, w/v) display identical shape. Finally, no

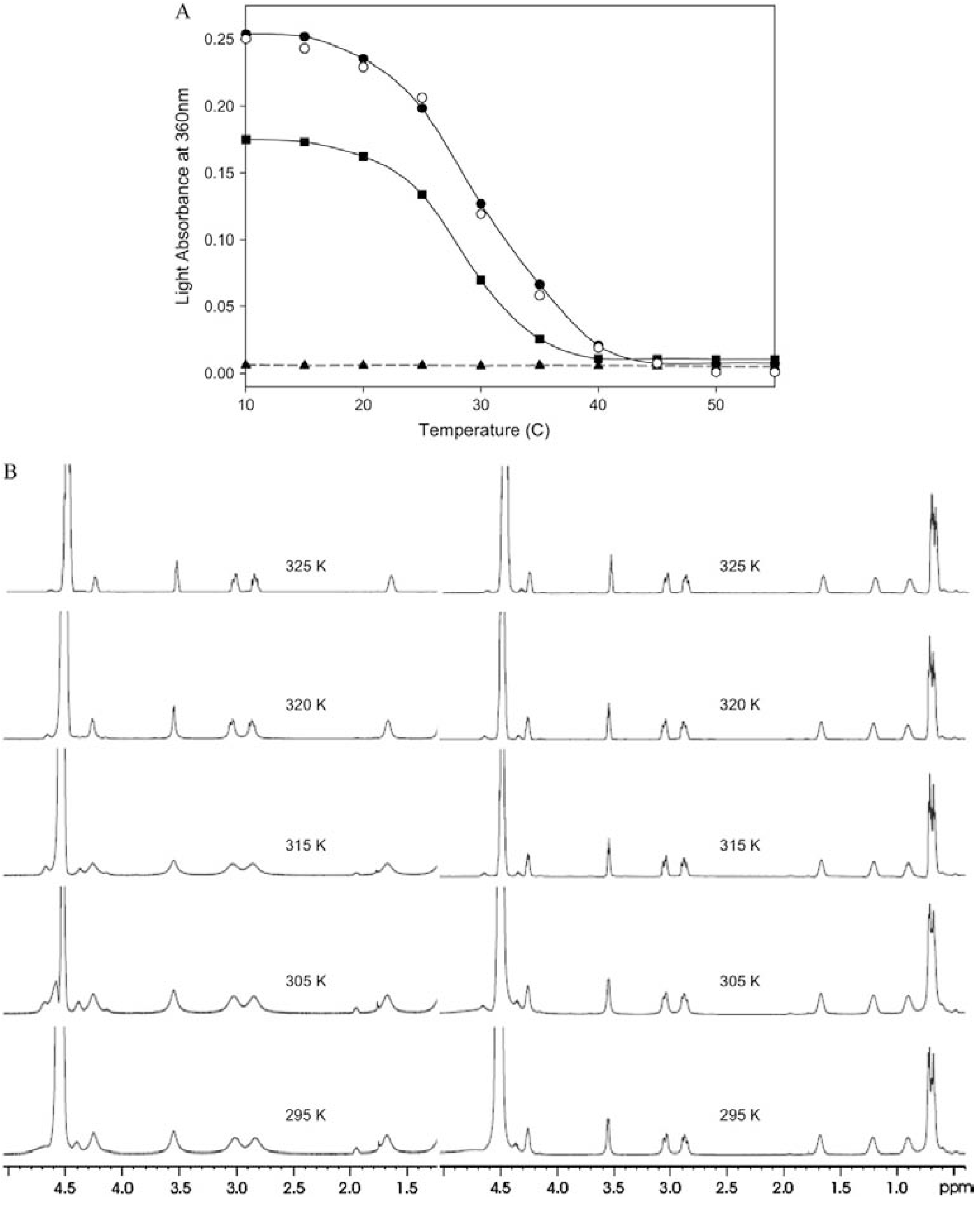


FIGURE 5 Dependence of peptide assembly state on the temperature. (*A*) Disassembly of 1.5% (w/v) (*solid squares*) and 2% (w/v) (*solid circles*) NH_3^+-Ile-Phe-COO^- dipeptide upon heating. Reassembly of 2% (w/v) NH_3^+-Ile-Phe-COO^-dipeptide upon cooling (*open circles*). No changes in absorbance were observed upon heating a 2% (w/v) NH_3^+-Val-Phe-COO^- sample (*solid triangles*). (*B*) ^{1}H NMR spectra of NH_3^+-Ile-Phe-COO^- (*left*) and NH_3^+-Val-Phe-COO^- (*right*) at increasing temperatures.

isosbestic point was observed that could reveal the transition between two spectroscopically different states of Phe when the absorption spectral changes were recorded for a 2% (w/v) sample at variable temperature (results not shown).

It has long been suggested that hydrophobic interactions play an important role in protein and peptide self-assembly (30). The presence of an additional methyl group in the Ile-Phe dipeptide relative to Val-Phe provides it with increased

TABLE 1 Selected FTIR bands of Ile-Phe and Val-Phe dipeptide samples at various concentrations and temperatures

	Concentration	Temperature		Amide	
Peptide	% (w/v)	K	NH	I	COO^-
Ile-Phe	0.1	298	3398	1662	1598
Ile-Phe	0.8	298	3396	1661	1598
Ile-Phe	2	298	3397	1660	1570
Ile-Phe	2	323	3387	1659	1597
Val-Phe	2	298	3396	1663	1598
Ile-Phe	2	298	3305	1618	1570
	dehydrated		3403	1658	
				1678	

Wave numbers are given in cm^{-1}.

hydrophobicity. We used the polarity-sensitive probe TNS to elucidate whether the self-association process is driven and/or stabilized by hydrophobic interactions. TNS binds with much higher affinity to surfaces or pockets formed by clusters of hydrophobic groups than to solvent-exposed isolated hydrophobic groups, resulting in an increase and blue-shift in the maximum of fluorescence emission compared with the emission of free TNS in aqueous solution. Little binding was detected for a 0.1% (w/v) sample of Ile-Phe (Fig. 6 *A*), confirming that the hydrophobic side chains of the dipeptides do not associate at low concentrations. In contrast, the probe binds strongly to the macromolecular structures formed in a 2% (w/v) sample, as proven by a large increase in fluorescence and a strong blue-shift of the maximum from 443 nm to 423 nm (Fig. 6 *A*), indicating the formation of a large hydrophobic environment upon self-assembly. The loss of most of the fluorescence signal upon heating the sample indicates the requirement of an ordered nanostructure for TNS to bind efficiently (Fig. 6 *A*). The process is again fully reversible and the TNS binding ability is restored upon cooling (data not shown). We took advantage of the reversibility of the process to study whether the self-assembly of the dipeptide and the formation of hydrophobic clusters occur simultaneously. A 2% (w/v) sample was heated to 333 K, and then progressively cooled down to 278 K, simultaneously monitoring the dipeptide self-assembly by light scattering and the formation of hydrophobic regions by TNS fluorescence emission. As can be observed in Fig. 6 *B*, the light-scattering dependence on the temperature is sigmoid and sharply corresponds to that reported by measuring absorbance at 360 nm. However, the increase in fluorescence emission occurs in two steps. The first, monotonic increase in fluorescence does not coincide with the light-scattering curve and may be interpreted as the hydrophobic interaction-governed self-assembly of the dipeptide in soluble oligomers. The concentration of these soluble assemblies saturates at ~321 K (after ~90s), as inferred from the difference between the TNS-binding and light-scattering signals (Fig. 6 *B*, *inset*). Accordingly, at these temperatures, NMR signals are well resolved, indicating high mobility of the building blocks and probably absence of rigid supramolecular structures. Several of these small aggregates would probably form a larger aggregate in a second or higher-order reaction and this aggregate would serve as the nucleus for the growth reaction visible from 150 s on, resulting in the rapid formation of larger assemblies detectable by light scattering, with a parallel increase in TNS binding and broadening of the NMR peaks. The data support a nucleation-growth pathway that gives rise to a remarkably high degree of cooperativity. This behavior is reminiscent of the formation of polypeptide aggregates, which usually exhibit a nucleated polymerization reaction in which an initial nucleation event is followed by the extension of newly formed nuclei into larger aggregates, including insoluble fibrils (31).

DISCUSSION

The dimensions of the fibrillar structures formed by the Ile-Phe dipeptide are similar to those reported for the nanotubes formed by diphenylalanine, the core recognition motif of Alzheimer's β-amyloid polypeptide (20). The crystal x-ray structure of the diphenylalanine peptide, as formed by fast evaporation of an aqueous solution of the peptide at high temperature, showed a crystal packing where the dipeptide forms aligned and elongated channels with a hydrophilic inner surface. The channels are lined with hydrogen-bond donors and acceptors in charged groups (NH_3^+ and COO^-), with hydrophobic side chains that act as a glue between the cylinders of peptide main chains and promote fiber formation (32,33). This model is compatible with the experimental data for the Ile-Phe fibrillar structures, including the observations that fibrillar structures are somehow transparent to electrons, the absence of β-sheet intermolecular contacts in the gel state, the observed NH_3^+/COO^- head-to-tail interactions, and the relevant role played by hydrophobicity in the assembly (Fig. 7).

Overall, the data herein allow us to propose a mechanism for the self-assembly of the Ile-Phe dipeptide into the observed nanostructures and to explain the self-association incompetence of the Val-Phe version. The initial establishment of intermolecular hydrogen bonds or electrostatic interactions between the extremes of dipeptide molecules or amide bonds is likely to be highly disfavored in water due to strong competitive solvent effects. At this stage, the establishment of intermolecular hydrophobic interactions between the side chains of Ile-Phe would drive the formation of sufficiently large primary soluble assemblies, which may act as nuclei or scaffolds for subsequent bonding. Such assemblies would further organize into nanofibrils by head-to-tail interactions between dipeptides. This hierarchical pathway for the self-assembly of Ile-Phe into fibrillar structures would explain the observed curves upon cooling a solution of molecularly dissolved monomers at high temperature. The fact that no specific FTIR signal, relative to that of the soluble monomer, could be detected at the temperature at which the small aggregates maximally populate (Fig. 6 *B* and Table 1) indicates that they more likely lack an ordered structure and

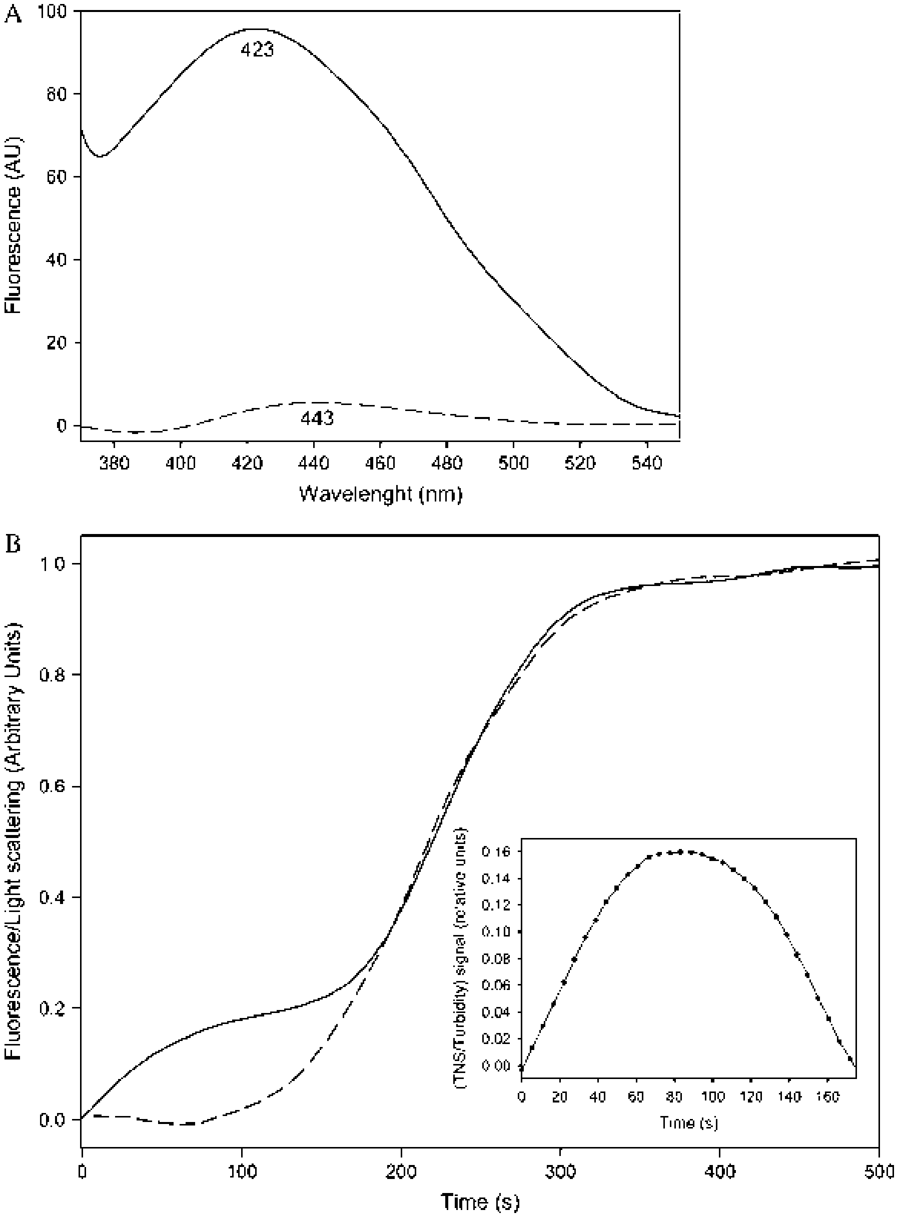


FIGURE 6 Role of hydrophobicity on dipeptide self assembly. (*A*) Fluorescence emission spectra of TNS incubated in the presence of 0.1% (w/v) (*dashed lines*) and 2% (w/v) (*solid lines*) NH_3^+-Ile-Phe-COO$^-$ dipeptide. (*B*) Kinetics of NH_3^+-Ile-Phe-COO$^-$ dipeptide self-assembly followed by light scattering (*dashed lines*) and TNS binding (*solid lines*). (*Inset*) Relative amounts of soluble self-assemblies found at the beginning of the polymerization process, as inferred from the difference between the TNS-binding and light-scattering signals.

specific bonding, pointing to hydrophobicity as the main driving force in the first stages of polymerization. It would be the subsequent establishment of specific, oriented, noncovalent contacts that might permit the formation of a highly ordered fibrillar superstructure. The lower hydrophobicity of the Val-Phe dipeptide would prevent the formation of the initial assemblies, thus aborting the subsequent nucleation and polymerization events. Alternatively, if the nucleation event occurs, but no specific contacts can be established thereafter, the result is the formation of nonordered amorphous aggregates probably stabilized by nonspecific hydrophobic interactions, as observed here for the Ile-Phe dipeptide at low or high pH.

The proposed dominating role of hydrophobic interactions at the beginning of the process may also explain the self-association properties of different dipeptides in the literature. For instance, whereas Phe-Phe has been shown to form nanotubes (20) and Phg-Phg spherical structures (20), the more polar Trp-Phe, Trp-Trp, and Trp-Tyr dipeptides were unable to self-assemble under the same conditions (20). These observations belie the role of aromatic stacking as a main assembly-driving force and point to the higher hydrophobicity of Phe and Phg as the mechanism responsible for the initial assembly reaction. According to this, Phe is the aromatic residue more commonly found in amyloid-forming peptides (34). Along with our data, and in contrast to previous assumptions, it has been shown that in the natural amyloid-forming peptide amylin the presence of an aromatic residue in the core is not necessary for amyloid formation and a large aliphatic residue performs equally well, whereas

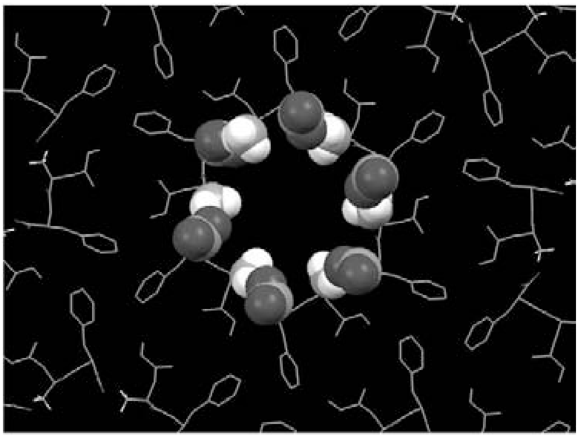


FIGURE 7 Molecular model of Ile-Phe self-assembled structures. The model is based on the crystal x-ray structure of the diphenylalanine peptide (33). Dipeptide backbone and hydrophobic side chains are shown as stick representations. NH_3^+ and COO^- terminal groups establishing head-to-tail interactions in the central channel are shown in spacefill representation.

substitution of the aromatic residue by an Ala results in very reduced aggregation (35). Also, a recent study of the aggregation of several mutants of human muscle acylphosphatase, in which aromatic residues were substituted with nonaromatic ones, shows that the changes in aggregation rates upon mutation arise predominantly from variations in hydrophobicity and intrinsic β-sheet propensity (36). Interestingly enough, the computational comparison of the binding propensities and the amyloid formation preferences of natural amino acids also revealed that Ile is the least structurally conserved residue in protein binding and at the same time has a high propensity for amyloid formation. This suggests that nature tends to avoid Ile conservation in protein-protein interactions to limit amyloid formation (37). Importantly, in the first study on the effects of mutation on the nucleation step of Aβ, it was shown that for position 18 of this peptide Ile is precisely the residue that promotes the fastest nucleation reaction (38). These observations are in full agreement with the significant correlations between aggregation and both hydrophobicity and β-sheet propensity that we found in the adjacent position 19 of the Aβ42 peptide (22), suggesting that Phe promotes aggregation because of these factors rather than for its aromaticity. Nevertheless, aromatic-aromatic interactions could still play a very important role in allowing specific contacts that dictate either the structure of the assembly, its stability, or the kinetics of self-assembly in peptide-derived nanostructures and amyloid fibrils.

Although the growth of fibrillar structures typically requires nucleation, the nature and properties of the nuclei and first soluble aggregates are still largely unknown due to the difficulty involved in characterizing them. The hydrophobic recruitment mechanism reported here could be of relevance to understanding the fibrillogenesis pathway of peptides, such us Aβ42. Recent studies show that Alzheimer's peptide fibril formation starts with the formation of globular amyloid-derived diffusible ligands (39) rather than with direct assembly of short protofibrils. Fibrils, and probably protofibrils, are stabilized by specific interactions (40), but if, as shown here for dipeptides, these interactions cannot efficiently trigger a self-assembly process in aqueous environment, it seems reasonable to propose that self-association begins with the formation of sufficiently large primary soluble globular structures (amyloid-derived diffusible ligands) driven by more or less unspecific hydrophobic contacts. As for dipeptides, their structural reorganization by specific interactions, including aromatic stacking, could turn them into short protofibrillar scaffolds instead of amorphous aggregates, with the ability to recruit and orientate new peptide units, acting as seeds of the fibrillogenic process. Understanding the details of the first steps of the aggregation/fibrillation mechanism at the molecular level is central to developing strategies for treatment or possible prevention of amyloid-deposition diseases. As shown here, in addition to their biotechnological applications, short peptides can also serve as ideal model systems for such studies.

We thank Dr. Esteve Padros, Dr. Josep Cladera, and their group for FTIR facilities.

This work was supported by grants BIO2004-05879 and CTQ2006-01080 from the Ministerio de Ciencia y Tecnología (MCYT) Spain, by the Centre de Referència en Biotecnologia, and grants SGR2005-00037 and SGR2005-01037 (Generalitat de Catalunya, Spain).

REFERENCES

1. Hartgerink, J. D., E. Beniash, and S. I. Stupp. 2001. Self-assembly and mineralization of peptide-amphiphile nanofibers. *Science*. 294:1684–1688.
2. Zhang, S. 2003. Fabrication of novel biomaterials through molecular self-assembly. *Nat. Biotechnol.* 21:1171–1178.
3. Du, C., G. Falini, S. Fermani, C. Abbott, and J. Moradian-Oldak. 2005. Supramolecular assembly of amelogenin nanospheres into birefringent microribbons. *Science*. 307:1450–1454.
4. Yoshimura, I., Y. Miyahara, N. Kasagi, H. Yamane, A. Ojida, and I. Hamachi. 2004. Molecular recognition in a supramolecular hydrogel to afford a semi-wet sensor chip. *J. Am. Chem. Soc.* 126:12204–12205.
5. Haines, L. A., K. Rajagopal, B. Ozbas, D. A. Salick, D. J. Pochan, and J. P. Schneider. 2005. Light-activated hydrogel formation via the triggered folding and self-assembly of a designed peptide. *J. Am. Chem. Soc.* 127:17025–17029.
6. Kisiday, J., M. Jin, B. Kurz, H. Hung, C. Semino, S. Zhang, and A. J. Grodzinsky. 2002. Self-assembling peptide hydrogel fosters chondrocyte extracellular matrix production and cell division: Implications for cartilage tissue repair. *Proc. Natl. Acad. Sci. USA*. 99:9996–10001.
7. Aggeli, A., M. Bell, N. Boden, J. N. Keen, P. F. Knowles, T. C. McLeish, M. Pitkeathly, and S. E. Radford. 1997. Responsive gels formed by the spontaneous self-assembly of peptides into polymeric β-sheet tapes. *Nature*. 386:259–262.
8. Wang, C., R. J. Stewart, and J. Kopecek. 1999. Hybrid hydrogels assembled from synthetic polymers and coiled-coil protein domains. *Nature*. 397:417–420.
9. Lyon, R. P., and W. M. Atkins. 2001. Self-assembly and gelation of oxidized glutathione in organic solvents. *J. Am. Chem. Soc.* 123:4408–4413.

10. Rajagopal, K., and J. P. Schneider. 2004. Self-assembling peptides and proteins for nanotechnological applications. *Curr. Opin. Struct. Biol.* 14:480–486.

11. Gazit, E. 2005. Mechanisms of amyloid fibril self-assembly and inhibition. Model short peptides as a key research tool. *FEBS J.* 272: 5971–5978.

12. Westermark, P. 2005. Aspects on human amyloid forms and their fibril polypeptides. *FEBS J.* 272:5942–5949.

13. Juszczak, L. J. 2004. Comparative vibrational spectroscopy of intracellular Tau and extracellular collagen I reveals parallels of gelation and fibrillar structure. *J. Biol. Chem.* 279:7395–7404.

14. Shen, C. L., M. C. Fitzgerald, and R. M. Murphy. 1994. Effect of acid predissolution on fibril size and fibril flexibility of synthetic β-amyloid peptide. *Biophys. J.* 67:1238–1246.

15. Esteras-Chopo, A., L. Serrano, and M. L. de la Paz. 2005. The amyloid stretch hypothesis: Recruiting proteins toward the dark side. *Proc. Natl. Acad. Sci. USA.* 102:16672–16677.

16. Ventura, S., J. Zurdo, S. Narayanan, M. Parreno, R. Mangues, B. Reif, F. Chiti, E. Giannoni, C. M. Dobson, F. X. Aviles, and L. Serrano. 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.

17. Ivanova, M. I., M. R. Sawaya, M. Gingery, A. Attinger, and D. Eisenberg. 2004. An amyloid-forming segment of β2-microglobulin suggests a molecular model for the fibril. *Proc. Natl. Acad. Sci. USA.* 101:10584–10589.

18. Madine, J., A. J. Doig, A. Kitmitto, and D. A. Middleton. 2005. Studies of the aggregation of an amyloidogenic alpha-synuclein peptide fragment. *Biochem. Soc. Trans.* 33:1113–1115.

19. Leffers, K. W., H. Wille, J. Stohr, E. Junger, S. B. Prusiner, and D. Riesner. 2005. Assembly of natural and recombinant prion protein into fibrils. *Biol. Chem.* 386:569–580.

20. Reches, M., and E. Gazit. 2003. Casting metal nanowires within discrete self-assembled peptide nanotubes. *Science.* 300:625–627.

21. Bitan, G., S. S. Vollers, and D. B. Teplow. 2003. Elucidation of primary structure elements controlling early amyloid β-protein oligomerization. *J. Biol. Chem.* 278:34882–34889.

22. de Groot, N. S., F. X. Aviles, J. Vendrell, and S. Ventura. 2006. Mutagenesis of the central hydrophobic cluster in Aβ42 Alzheimer's peptide. Side-chain properties correlate with aggregation propensities. *FEBS J.* 273:658–668.

23. Zurdo, J., J. I. Guijarro, J. L. Jimenez, H. R. Saibil, and C. M. Dobson. 2001. Dependence on solution conditions of aggregation and amyloid formation by an SH3 domain. *J. Mol. Biol.* 311:325–340.

24. Muni, N. J., H. Qian, N. M. Qtaishat, R. A. Gemeinhart, and D. R. Pepperberg. 2006. Activation of membrane receptors by neurotransmitter released from temperature-sensitive hydrogels. *J. Neurosci. Methods.* 151:97–105.

25. Hokugo, A., M. Ozeki, O. Kawakami, K. Sugimoto, K. Mushimoto, S. Morita, and Y. Tabata. 2005. Augmented bone regeneration activity of platelet-rich plasma by biodegradable gelatin hydrogel. *Tissue Eng.* 11:1224–1233.

26. Venyaminov, S., and N. N. Kalnin. 1990. Quantitative IR spectrophotometry of peptide compounds in water (H2O) solutions. I. Spectral parameters of amino acid residue absorption bands. *Biopolymers.* 30:1243–1257.

27. Kalnin, N. N., I. A. Baikalov, and S. Venyaminov. 1990. Quantitative IR spectrophotometry of peptide compounds in water (H2O) solutions. III. Estimation of the protein secondary structure. *Biopolymers.* 30: 1273–1280.

28. Venyaminov, S., and N. N. Kalnin. 1990. Quantitative IR spectrophotometry of peptide compounds in water (H2O) solutions. II. Amide absorption bands of polypeptides and fibrous proteins in α-, β-, and random coil conformations. *Biopolymers.* 30:1259–1271.

29. Eker, F., X. Cao, L. Nafie, and R. Schweitzer-Stenner. 2002. Tripeptides adopt stable structures in water. A combined polarized visible Raman, FTIR, and VCD spectroscopy study. *J. Am. Chem. Soc.* 124: 14330–14341.

30. Deechongkit, S., E. T. Powers, S. L. You, and J. W. Kelly. 2005. Controlling the morphology of cross β-sheet assemblies by rational design. *J. Am. Chem. Soc.* 127:8562–8570.

31. Thirumalai, D., D. K. Klimov, and R. I. Dima. 2003. Emerging ideas on the molecular basis of protein and peptide aggregation. *Curr. Opin. Struct. Biol.* 13:146–159.

32. Gorbitz, C. H. 2001. Nanotube formation by hydrophobic dipeptides. *Chemistry (Easton).* 7:5153–5159.

33. Gorbitz, C. H. 2006. The structure of nanotubes formed by diphenylalanine, the core recognition motif of Alzheimer's β-amyloid polypeptide. *Chem. Commun.* (*Camb.*). 2332–2334.

34. Gazit, E. 2002. A possible role for pi-stacking in the self-assembly of amyloid fibrils. *FASEB J.* 16:77–83.

35. Tracz, S. M., A. Abedini, M. Driscoll, and D. P. Raleigh. 2004. Role of aromatic interactions in amyloid formation by peptides derived from human Amylin. *Biochemistry.* 43:15901–15908.

36. Bemporad, F., N. Taddei, M. Stefani, and F. Chiti. 2006. Assessing the role of aromatic residues in the amyloid aggregation of human muscle acylphosphatase. *Protein Sci.* 15:862–870.

37. Ma, B., and R. Nussinov. 2006. Trp/Met/Phe hot spots in protein-protein interactions: potential targets in drug design. *Curr. Top. Med. Chem.* In press.

38. Christopeit, T., P. Hortschansky, V. Schroeckh, K. Guhrs, G. Zandomeneghi, and M. Fandrich. 2005. Mutagenic analysis of the nucleation propensity of oxidized Alzheimer's β-amyloid peptide. *Protein Sci.* 14:2125–2131.

39. Lambert, M. P., A. K. Barlow, B. A. Chromy, C. Edwards, R. Freed, M. Liosatos, T. E. Morgan, I. Rozovsky, B. Trommer, K. L. Viola, P. Wals, C. Zhang, C. E. Finch, G. A. Krafft, and W. L. Klein. 1998. Diffusible, nonfibrillar ligands derived from Aβ1–42 are potent central nervous system neurotoxins. *Proc. Natl. Acad. Sci. USA.* 95: 6448–6453.

40. Petkova, A. T., Y. Ishii, J. J. Balbach, O. N. Antzutkin, R. D. Leapman, F. Delaglio, and R. Tycko. 2002. A structural model for Alzheimer's β-amyloid fibrils based on experimental constraints from solid state NMR. *Proc. Natl. Acad. Sci. USA.* 99:16742–16747.

BMC Structural Biology

Research article

Prediction of "hot spots" of aggregation in disease-linked polypeptides

Natalia Sánchez de Groot†[1], Irantzu Pallarés†[1], Francesc X Avilés[1,2], Josep Vendrell[1,2] and Salvador Ventura*[1,2]

Address: [1]Departament de Bioquímica i Biologia Molecular, Facultat de Ciències, Universitat Autònoma de Barcelona, E-08193 Bellaterra, Spain and [2]Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, E-08193 Bellaterra, Spain

Email: Natalia Sánchez de Groot - Natalia.Sanchezd@campus.uab.es; Irantzu Pallarés - irantzu.pallares@uab.es; Francesc X Avilés - FrancescXavier.Aviles@uab.es; Josep Vendrell - josep.vendrell@uab.es; Salvador Ventura* - salvador.ventura@uab.es

* Corresponding author   †Equal contributors

Published: 30 September 2005

BMC Structural Biology 2005, **5**:18   doi:10.1186/1472-6807-5-18

Received: 28 September 2005
Accepted: 30 September 2005

This article is available from: http://www.biomedcentral.com/1472-6807/5/18

Abstract

Background: The polypeptides involved in amyloidogenesis may be globular proteins with a defined 3D-structure or natively unfolded proteins. The first class includes polypeptides such as β2-microglobulin, lysozyme, transthyretin or the prion protein, whereas β-amyloid peptide, amylin or α-synuclein all belong to the second class. Recent studies suggest that specific regions in the proteins act as "hot spots" driving aggregation. This should be especially relevant for natively unfolded proteins or unfolded states of globular proteins as they lack significant secondary and tertiary structure and specific intra-chain interactions that can mask these aggregation-prone regions. Prediction of such sequence stretches is important since they are potential therapeutic targets.

Results: In this study we exploited the experimental data obtained in an *in vivo* system using β-amyloid peptide as a model to derive the individual aggregation propensities of natural amino acids. These data are used to generate aggregation profiles for different disease-related polypeptides. The approach detects the presence of "hot spots" which have been already validated experimentally in the literature and provides insights into the effect of disease-linked mutations in these polypeptides.

Conclusion: The proposed method might become a useful tool for the future development of sequence-targeted anti-aggregation pharmaceuticals.

Background

In the last decade, protein aggregation has moved beyond being a mostly ignored area of protein chemistry to become a key topic in medical sciences [1], mainly because the presence of insoluble deposits in human tissues correlates with the development of many debilitating human disorders including the amyloidoses and several neurodegenerative diseases [2]. The proteins involved in these diseases are not related in terms of sequence or secondary structure content. From the conformational point of view, two major classes can be distinguished: globular proteins with a stable unique conformation in the native state and intrinsically unstructured proteins [3]. Globular proteins rarely aggregate from their native states and destabilization, resulting in an increased population of unfolded molecules, is well established as a trigging factor in disorders associated with the deposition of proteins that are globular in their normal functional states [4], as in the cases of β2-microglobulin, lysozyme, transthyretin and the prion protein. Interestingly enough, many proteins involved in depositional disorders are mostly unstructured within the cell [3]. These include amylin, amyloid-β-protein, and α-synuclein, among others. In these cases, protein deposition does not require unfolding and can occur by direct self-assembly of the unstructured polypeptide chains.

One of the major unanswered questions of protein aggregation is the specificity with which the primary sequence determines the aggregation propensity from totally or partially unfolded states. Deciphering the answer to this question will give us a chance to control the unwanted protein deposition events through specific sequence-targeted therapeutics. A first advance in this direction is the recent discovery that not all regions of a polypeptide are equally important for determining its aggregation tendency, both in natively unfolded and globular proteins. In this way, some authors, including ourselves, have proved recently that very short specific amino acid stretches can act as facilitators or inhibitors of amyloid fibril formation [5,6]. These relevant regions are usually known as aggregation "hot spots". Aggregation-prone regions are likely to be blocked in the native state of globular proteins because their side chains are usually hidden in the inner hydrophobic core or already involved in the network of contacts that stabilizes a protein. This accounts for the protective role of the native structure against aggregation [7]. In contrast, aggregation-prone regions are already exposed to solvent in natively unfolded proteins, available for the establishment of inter-molecular contacts that may finally lead to the formation of aggregates. Accordingly, the presence of putative "hot spots" of aggregation is much more frequent in the sequences of globular proteins than in those coding for natively unfolded proteins [8]. The presence of aggregation-prone regions has been described in most of the peptides and proteins underlying neurodegenerative and systemic amyloidogenic disorders [9].

We have used a simple *in vivo* system to study the aggregation effects of a complete set of mutations in one of the best characterized "hot spots" in a disease-linked protein: the central hydrophobic cluster (CHC) of the Amyloid-β-protein (Aβ) [10,11]. The results in this and other studies on protein models not related to disease [12], suggested that common and simple principles underlie protein aggregation, at least from totally or partially unfolded states, and that the propensities of proteins backbones to aggregate are sharply modulated by the sequences that dress them. Based on these assumptions, we have developed a simple approach that identifies the presence of "hot-spots" of aggregation in globular and unstructured disease-linked polypeptides and predicts the aggregation effects of mutations in their sequences.

Results and discussion

Aggregation propensities of natural amino acids

The rationale behind our study is based on two recent observations in the field. First, not all the polypeptide sequence is relevant for the aggregation of a given protein, but rather there exist specific regions that drive the process [5,6] and second, similar simple rules appear to underlie the aggregation propensities of unrelated proteins from unfolded states [12]. According to these two assumptions one may expect that the conclusions obtained from the study of a relevant "hot spot" of aggregation in a specific protein could apply to other unrelated proteins involved in disease. As commented upon previously, we have exploited an *in vivo* reporter method to calculate the relative aggregation propensities of each individual natural amino acid when placed in the central position of the CHC of Aβ (see Material and Methods). The highest aggregation propensities correspond to isoleucine, phenylalanine, valine, and leucine, whereas aspartic, glutamic, asparagine, and arginine exhibit the lowest (Table 1). In general, hydrophobic residues tend to induce aggregation whereas polar ones promote solubility, matching the general assumption that hydrophobic interactions are supposed to play an important role in protein aggregation [13].

Table 1: Relative experimental aggregation propensities of the 20 natural amino acids derived from the analysis of mutants in the central position of the CHC in amyloid-β-protein.

Amino acid	
I	1.822
F	1.754
V	1.594
L	1.380
Y	1.159
W	1.037
M	0.910
C	0.604
A	-0.036
T	-0.159
S	-0.294
P	-0.334
G	-0.535
K	-0.931
H	-1.033
Q	-1.231
R	-1.240
N	-1.302
E	-1.412
D	-1.836

Generation of protein aggregation profiles and prediction of the effects of protein mutation on the aggregation propensity

Provided that a given polypeptide aggregates from an at least partially unstructured state, the experimental intrinsic aggregation propensities shown in Table 1 should apply independently of the protein context. Thus, a profile can be theoretically generated for any protein sequence to detect those regions with aggregation propensities above the average value of the whole sequence. This leads directly to the definition of "hot spot" of aggregation as a certain region that displays higher aggregation propensity than the rest of the sequence. Interestingly, a related approach has been reported very recently for the analysis of unstructured proteins associated with neurodegenerative diseases[14].

A good number of natural occurring mutations have been reported in proteins associated to depositional diseases. In many cases they result in changes in the global protein aggregation propensity and sometimes in the appearance of premature or acute pathological symptoms. The change in average aggregation propensity (ΔAP) between the wild type and the different mutants should predict the effect of sequence variations on the aggregation propensities, provided that they rely on changes in the intrinsic polypeptide properties.

Analysis of disease-related polypeptide sequences

In this section the above described analysis is applied to a set of proteins linked to depositional diseases and the obtained results are compared with the available experimental data.

Intrinsically unstructured proteins

Amyloid-β-protein

As a proof of principle our approach was first tested in the molecule from which the experimental amino acid aggregation propensities were derived. Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by the patient's memory loss and impairment of cognitive abilities. The extracellular amyloid is found in the brain and is widely believed to be involved in the progression of the disease [15]. The principal component of the lesions is the hydrophobic polypeptide Aβ. The most abundant forms found in amyloid plaques are a 40-mer (Aβ40) and a 42-mer (Aβ42). Although less abundant, Aβ42 is more amyloidogenic than Aβ40 and is the major component of neuritic plaques [16]. Two main regions with high aggregation propensity can be distinguished in the aggregation profile for this polypeptide (Fig. 1). The second region arises from the contribution of two sequence stretches comprising residues 30–36 and 38–42, respectively. The predicted aggregation-prone regions are in excellent agreement with the experimental data in the literature. Residues 16–21 overlap with the CHC sequence comprising residues 17–21, a particular region recognized to play a key role in Aβ aggregation and that is defined as specially relevant for the amyloidogenesis of the Aβ40 and Aβ42 peptides by two recent proline-scanning-mutagenesis studies [17,18]. In addition, structural studies using solid state-NMR [19] and site-directed spin labeling [20] have revealed that residues 16–21 are located in the core of Aβ fibrils. Accordingly, a short 7 residues fragment comprising residues 16–22 is able to form ordered amyloid fibrils [21] and, more interestingly, 16-LVAFF-20 and derived peptides have been shown to bind to Aβ42 and act as potent inhibitors of amyloid formation [22]. The region 30–42, including both 30–36 and 38–42 stretches, has been also implicated in Aβ aggregation. Proline-scanning-mutagenesis revealed that the region 31–36 is sensitive to proline replacement and likely to include a β-sheet portion of the Aβ fibrils [17,18]. The contribution of the C-terminal region 38–42 to Aβ amyloidogenesis becomes clear from the observation that, although Aβ40 is produced in greater abundance *in vivo*, the prevalence of the full-length 42-mer in plaques is much higher [16]. Experiments with truncated synthetic Aβ peptides have confirmed that Aβ39 and Aβ40 are kinetically soluble for several days, whereas Aβ42 immediately aggregates into amyloid fibrils [23]. The relevance of the predicted 30–42 region is confirmed by structural studies that demonstrate that residues 30–40 are located in the core of the Aβ fibrils [20].

A set of mutations in the CHC and adjacent positions of Aβ42 is intimately associated to early-onset familial Alzheimer diseases (FAD). The substitutions include A21G (Flemish), E22Q (Dutch) and E22G (Arctic) [24]. Aβ42 congeners bearing these mutations display distinct aggregation kinetics. The rate of fibril formation by the Flemish mutant is decreased relative to WT Aβ42, whereas the Dutch mutant peptide aggregates substantially faster. The Arctic peptide does not shows an overall change in the rate of fibrillogenesis relative to WT Aβ42, but rather accelerated protofibril formation. To assess whether the effect of such mutations could be predicted by the present approach we calculated ΔAP for the different sequences. The results obtained describe accurately the effects documented in the literature (Table 2).

Adding to the mutations present in the population, a large set of mutations has been artificially introduced on Aβ that result in changes in its aggregation propensity. ΔAP values were also calculated for several of them and the results compared with the experimental data (Table 2). The calculated changes in aggregation propensity are in excellent agreement with the trends reported in the literature. Briefly, we predict the changes in aggregation of F19 mutants, those of I31 and I32 in the 30–36 region and

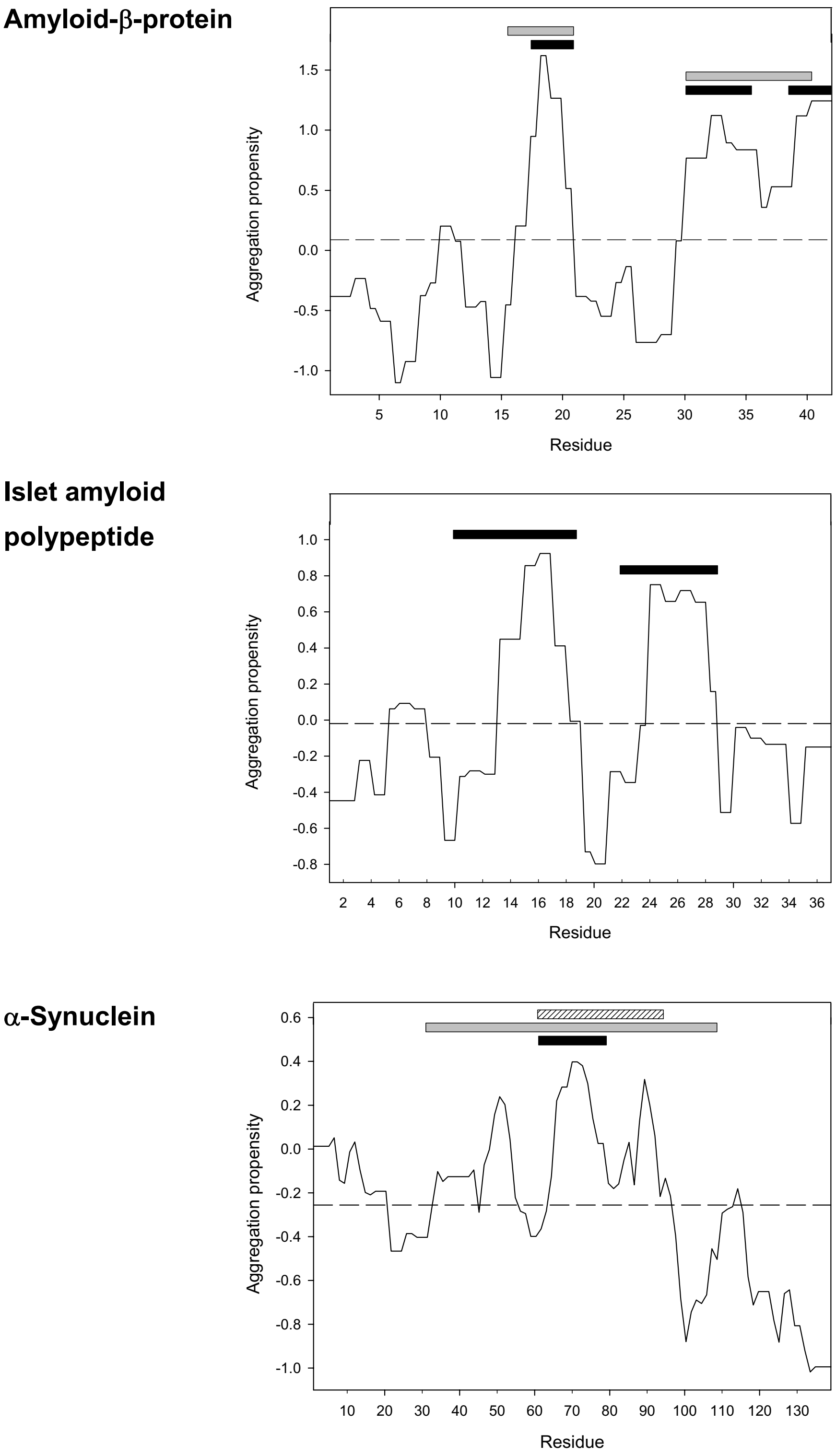


Figure 1

Aggregation profile of natively unfolded proteins related to disease. The average aggregation propensity of the different polypeptides is shown as a dashed line. Minimal protein regions which have been experimentally proven to be involved in aggregation are shown at the top of the plot as black bars. Regions in the core of the fibrils are shown as grey bars (when information available). The NAC fragment of α-Synuclein is shown as a dashed bar.

Table 2: Comparison of predicted and experimental changes in aggregation for Aβ variants.

Mutation	ΔAP*	Observed aggregation‡
A21G	-1.22	-
E22G	+2.14	+
E22Q	+0.44	+
F19P	-5.09	-
F19T	-4.73	-
I31L	-1.07	-
I32L	-1.07	-
I41G	-5.76	-
I41A	-4.52	-
I41L	-1.075	-
A42G	-1.21	-
A42V	+3.95	+
Δ1–4	+4.82	+
Δ1–9	+21.86	+
Δ40–42	-8.34	-
Δ41–42	-4.26	-
V12E+V18E+M35T+I41N	-18.96	-
F19S+L34P	-9.18	-

* Change in average aggregation propensity
‡ Changes in aggregation determined experimentally.

those of I41 and A42 in the C-terminal region, as well as the effects of deletions both in the N and the C ends. Finally, we also predict the high solubility of Aβ versions generated by random mutagenesis [25].

Islet amyloid polypeptide

Type II diabetes is associated with progressive beta-cell failure manifested as a decline in insulin secretion and increasing hyperglycemia. A growing body of evidence suggests that beta-cell failure in type II diabetes correlates with the formation of pancreatic islet amyloid. Islet amyloid polypeptide (IAPP, amylin), the major component of islet amyloid, is co-secreted with insulin from beta-cells. In type II diabetes, this peptide aggregates to form amyloid fibrils that are toxic to beta-cells [26]. IAPP is an unstructured peptide hormone of 37 amino acid residues. Two "hot spots" of aggregation comprising residues 12–18 and 22–28 are detected for this peptide (Fig. 1). Interestingly enough, a 8–37 IAPP-fragment including both "hot spots", has been shown to form amyloid fibrils under physiological conditions [27]. The two aggregation prone regions sharply coincide with those protected in the core of the fibrils in a recently described structural model of IAPP aggregates [28]. In this study, residues 12–17 and 22–27 are proposed to form the inner β-sheets in the fibril protofilament structure. According to this hypothesis, peptides corresponding to residues 8–20, 10–19, 20–29 of human IAPP, which include one of the "hot spots" described here, all form amyloid [29-31]. Smaller peptides derived from these regions have also been shown to form amyloid, and a recent investigation suggests that the minimal amyloid forming fragment of IAPP consists of residues 22–27. This hexapeptide fragment, NFGAIL, forms β-sheet-containing fibrils that coil around each other in typical amyloid fibril morphology [32].

The analysis also explains the available mutational data on IAPP. Diabetes-associated IAPP amyloid occurs in primates and cats but not in rodents [33]. Consistently, the sequences of peptides 20–29 of rodents display reduced average aggregation propensity relative to that of cat and human (Table 3). We also predict the slightly increased aggregation propensities of single or multiple mutations of rat IAPP to the corresponding residues of human IAPP [33]: R18H, L23F or V26I, as well as the results from alanine-scanning-mutagenesis in a peptide encompassing residues 22–27 [32] (Table 3). It has been found that a substitution at position 20 (S20G) in the IAPP molecule in a reduced subpopulation of Japanese people with type II diabetes is associated with an earlier onset and more severe form of disease [34]. In this case, our approach does no predict an increased but a slightly reduced aggregation propensity in the mutant, suggesting that the pathological symptoms in this variant may arise from non-intrinsic factors. In fact, it has been suggested that the accelerated aggregation of the S20G variant could be related to structural reasons, resulting from a better packing of the turns connecting the β-sheets in the final protofilament structure [28] that cannot be predicted by the present approach.

Several mechanisms have been proposed for IAPP fibril formation in type II diabetes. One widely accepted mechanism is that in type II diabetes, increased production and secretion of IAPP associated with increased demand for insulin might result in accumulation and aggregation of IAPP [35]. A second view considers that impaired processing of the IAPP precursor molecule, proIAPP, by islet beta-cells may lead to hypersecretion of unprocessed or partially processed forms of proIAPP that may have a higher tendency for aggregation compared to mature IAPP [35]. Our calculated average aggregation propensities for proIAPP and processed IAPP support this view (Table 3).

α-Synuclein

Parkinson disease is the most common neurodegenerative movement disorder and is pathologically characterized by the presence of neuronal intracytoplasmatic deposits of aggregated protein called Lewy bodies [36]. Lewy bodies also occur in other cognitive disorders, globally known as α-synucleinopathies. α-Synuclein is the major component of the fibrils that form the Lewy bodies [36]. It is a small (137 residues), natively unfolded, soluble, presynaptic and highly conserved protein without a well-defined

Table 3: Comparison of predicted and experimental changes in aggregation for IAPP variants, relative to the corresponding human IAPP sequence.

Variant	ΔAP*	Observed aggregation‡
(20–29) Cat	-5.12	=
(20–29) Rat	-16.46	-
(20–29) Hamster	-32.73	-
R18H	+0.94	+
L23F	+1.70	+
V26I	+0.42	+
R18H+L23F+V26I	+3.06	+
(22–27) N22A	+31.53	+
(22–27) F23A	-42.96	-
(22–27) G24A	+11.96	+
(22–27) I26A	-44.5872	-
(22–27) L27A	-33.99	-
S20G	-1.09	+
ProIAPP	+31.40	+?

* Change in average aggregation propensity
‡ Changes in aggregation determined experimentally.
? Not yet proved experimentally.

function. The aggregation profile for this polypeptide is shown in Fig. 1. Several large aggregation-prone stretches were predicted for the α-Synuclein sequence: region 1–18, region 27–56 and specially region 61–94. Again, our predictions are in complete agreement with the experimental data in the literature, as many studies suggest that the central region of the protein, known as the non-Aβ component of amyloid plaques (NAC, amino acids 61–95), is the responsible for its aggregation process [37]. A peptide comprising residues 68–78 of α-synuclein has been shown to be the minimum fragment that, like α-synuclein itself, forms amyloid fibrils and exhibits toxicity towards cells in culture [38]. This fragment is included in the region 62–80 which we predict as the sequence stretch with the highest aggregation propensity. All the α-synucleinopathies are characterized by the accumulation of the 35 residues NAC fragment in the insoluble deposits [37]. Accordingly, this central region is predicted to have a much higher average aggregation propensity than its soluble precursor (ΔAP = +38.47). The importance of this hydrophobic stretch is further supported by its absence in β-synuclein, a homologue of α-synuclein, with strongly reduced propensity for fibril formation. It has been shown that the deletion of amino acids 71–82 within the hydrophobic region abrogated the ability of human α-synuclein to polymerize into fibrils [39]. Protease digestion studies suggest that the core region of α-synuclein in the fibrils could be longer, since a 7-kDa fragment (comprising residues 31–109) was shown to be protected from proteinase K digestion [40]. This region contains the putative 12-residue core domain, as well as the NAC region and includes the second and third "hot spots" in our profile. A structural study on the organization of α-synuclein in the fibrilar state using site-directed spin labelling confirms that the 34–101 residues region constitutes the core of the fibrils forming a parallel in-register β-sheet structure whereas the N terminus is structurally more heterogeneous and the C terminus (40 amino acids) is completely unfolded [41].

Several α-Synuclein mutations appear associated with familial early-onset Parkinson Disease: A30P, A53T and E46K. All they map into our predicted second "hot spot". The rates of fibril assembly of the E46K and A53T mutants have been shown to be greater than those of the wild type and A30P proteins [42]. We predict a similar average aggregation propensity for the wild-type and the A30P mutant and an slightly increased aggregation propensity for the E46K mutant, but fail to foresee the effect of the A53T mutation in promoting the formation of protofibrils. Obviously, other functional factors apart from the intrinsic aggregation propensities can strongly influence the aggregation tendency of unfolded polypeptide chains within the cell. In fact the effects of α-synuclein mutations have been associated either to an impaired degradation inside lysosomes or to a reduced axonal transport of the variants [43,44]. Both situations may result in increased concentrations of the protein in certain regions of the neuron that may favor the nucleation step of amyloid formation. According to this, α-synuclein gene triplication identified in two independent families [45] has been shown to accelerate the development of Parkinson disease. Thus, an increase in the amount of cellular α-synuclein appears to be important for the pathogenesis of Parkinson disease, suggesting that the effects of the different α-synuclein mutations on protein aggregation could be quantitative, in terms of local concentration, rather that qualitative. Thus, experimental deviations from the theoretical predictions in natively unfolded proteins, in addition to reflect limitations of the approach, might also contain relevant information, prompting to find alternative structural, as in the case of amylin, or functional, as in the case of α-synuclein, explanations for the observed behavior.

Globular proteins

β2-Microglobulin

β2-Microglobulin-related amyloidosis is a common and serious complication in patients on longterm hemodialysis [46]. Intact β2-microglobulin is a major structural component of the amyloid fibrils. β2-Microglobulin (β2-m) is a small (99 residues) non-glycosilated protein with an immunoglobulin-like fold consisting in two antiparallel pleated β-sheets linked by a disulfide bond (Fig. 2). β2-m has been shown to form amyloid fibrils *in vitro* under different conditions, but in all cases β2-m populates

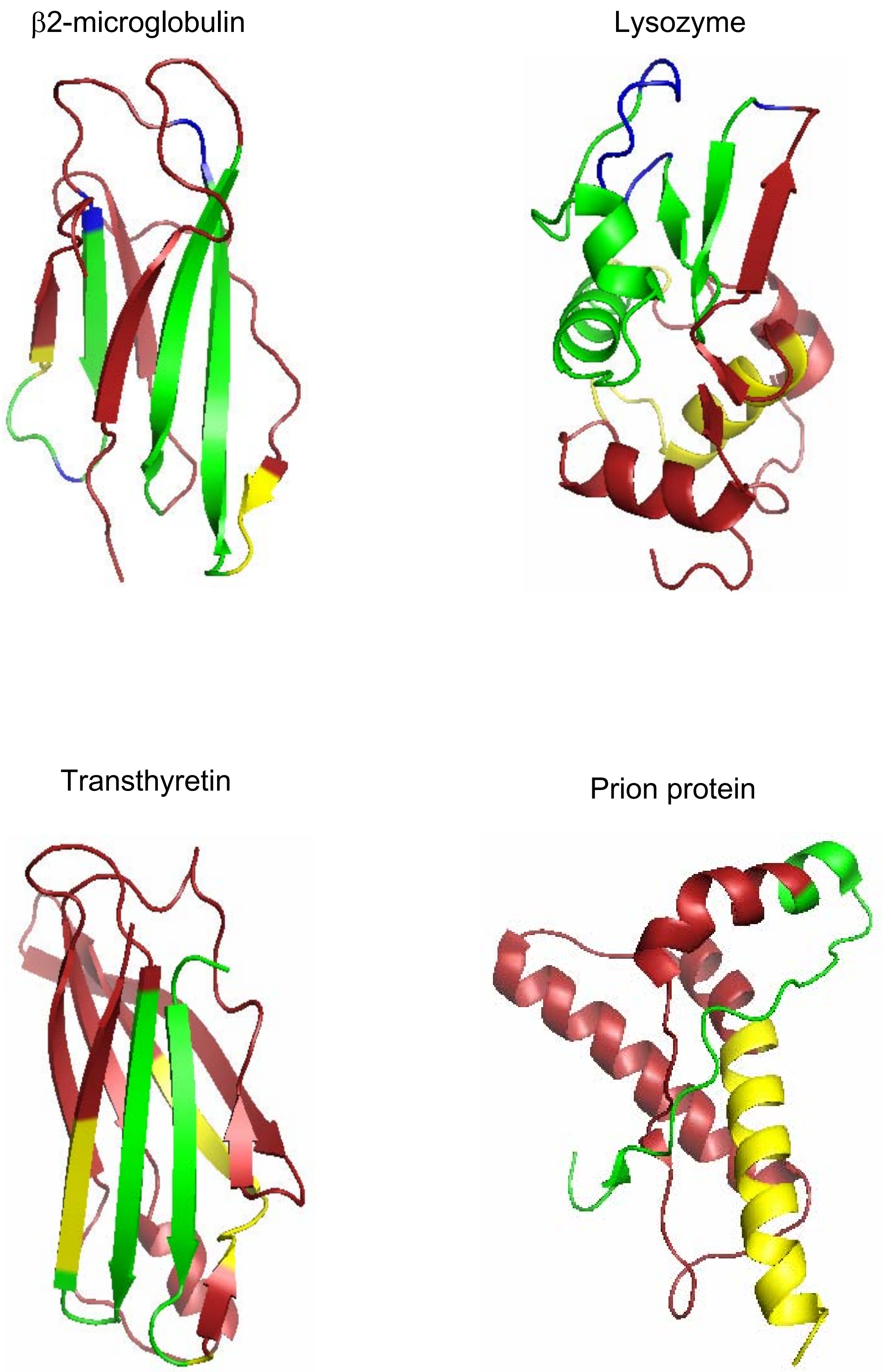


Figure 2

Representation of the 3D structure of globular proteins related to disease. The chain segments in which the prediction and the experimental data coincide are colored in green. Those identified experimentally to be relevant for amyloid formation but not predicted by the present approach are colored in blue. The regions predicted to be important for amyloid formation from which experimental data are not available or indicates that they are not involved in aggregation are shown in yellow.

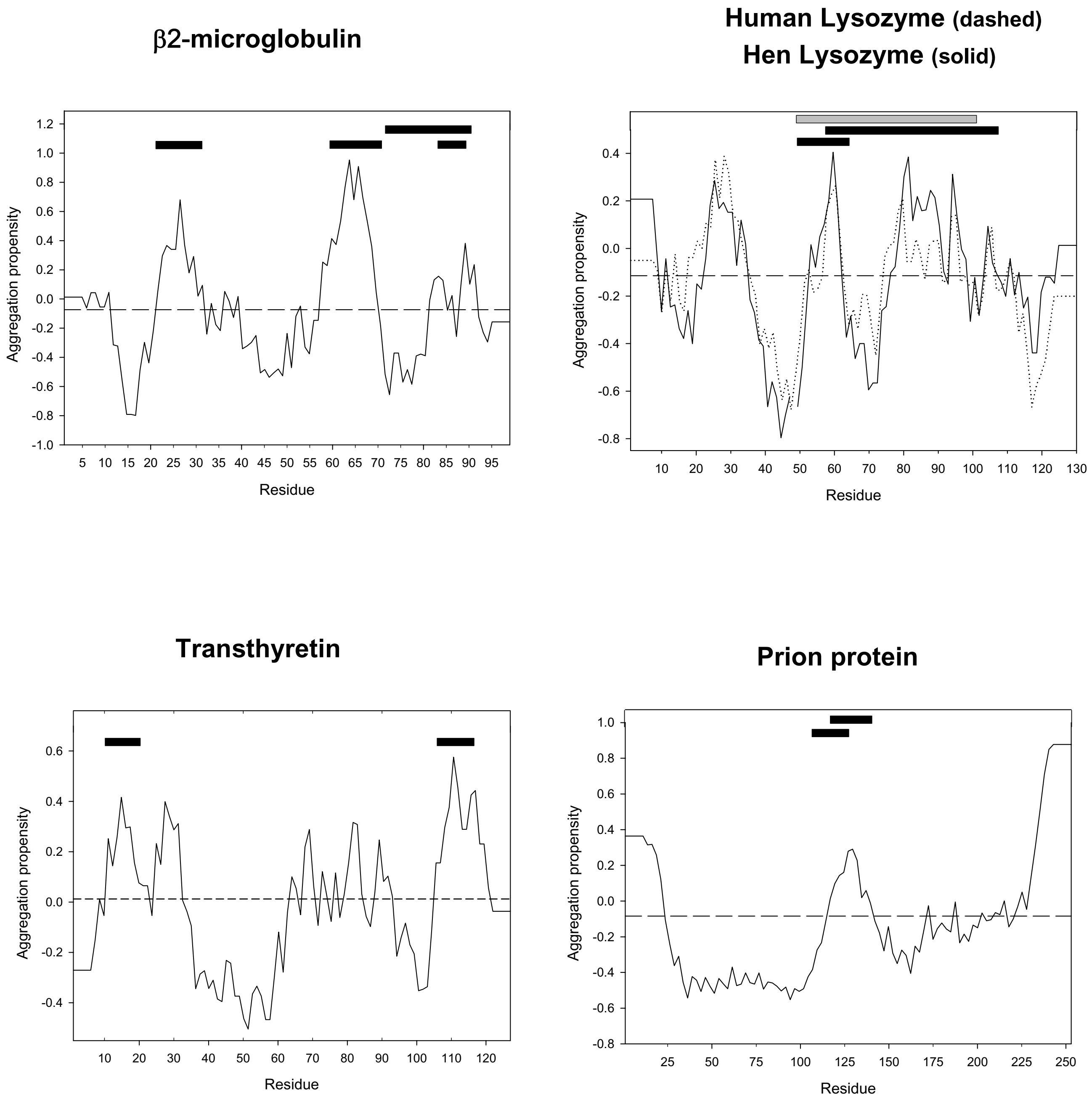


Figure 3

Aggregation profile of globular proteins related to disease. Minimal protein regions which have been proved experimentally to be involved in aggregation are shown at the top of the plot as black bars. Regions in the core of the fibrils are shown as a grey bars (when information available).

unfolded non-native states as precursors to fibril assembly [47]. Under these conditions aggregation-prone regions, if present, may promote and drive the aggregation event. According to the analysis of the aggregation profile, shown in Fig. 3, this protein displays four "hot spots" encompassing residues 21–31, 56–69 and 79–85, and 87–91. These regions sharply coincide with four different secondary structure elements in β2-m: β-strand 2, formed by residues 21–31; β-strand 6, formed by residues 61–71; β-strand 7; formed by residues 77–85 and β-strand 8, formed by residues 86–95 (Fig. 2). In agreement with our prediction a peptide comprising residues 21–41 has been shown to form fibrils in isolation [48]. In addition, a N-terminal fragment of this short peptide corresponding exactly to our first "hot spot" [21-31] is also able to self-assemble into fibrillar structures [49]. Interestingly enough, the peptides 23–31 and 21–29 exhibited reduced amyloidogenesis [49]. Thus, in this particular "hot spot" the prediction delimits not only the overall region important for aggregation but also its precise size. The amino acid stretches 59–79 and its shorter version 59–71 which overlap with the predicted second aggregation-prone region of β2-m have been also shown to form fibrils [50]. The C-terminal fragment 72–99 of β2-m has been also reported to form amyloid [51]. This 29 residues sequence includes our third and fourth "hot spots" of aggregation. The peptide 91–99 does not aggregate, indicating that the last 9 residues of β2-m are not relevant for amyloidogenesis as predicted here [49]. The N-terminal region, for which no aggregation propensity is predicted, is probably not involved inthe aggregation process as evidenced by the fact that the fragment 6–12 does not form fibrils [49]. This observation could be physiologically relevant since the N-terminus of β2-m is truncated in 30% of the molecules extracted from ex vivo fibrils [52].

In contrast to the human protein, mouse β2-m does not form fibrils even at high concentration [53]. Based on this observation a seven residues region corresponding to residues 83–89 of human β2-m has been suggested to be particularly important for aggregation, since it corresponds to the sequence with the highest divergence between both species. This hypothesis has been tested experimentally, since a heptapeptide bearing the human sequence is able to self-assemble whereas the mouse version is not [53]. The complete mouse sequence is predicted to have a strongly reduced aggregation propensity (ΔAP = -47.86).

Overall, our predictions on the presence and location of "hot spots" in β2-m are extremely accurate and overlap with the experimentally found relevant regions (Fig. 2). The observation that short peptides including the aggregation-prone regions described here form amyloids implies that exposure of previously hidden short segments can nucleate native proteins into the amyloid state and reinforces the hypothesis that fibril formation is sequence specific.

One of the most urgent issues in the study of amyloid fibrils is to reproduce the formation of fibrils under physiological conditions. Recently, it has been found that low concentrations of SDS around the critical micelle concentration induce the extensive growth of β2-m amyloid fibrils at physiological pH, probably through the SDS-induced conformational change of β2-m monomers [54]. Contrarily to what was expected, the presence of low concentration of SDS had little effect on the stability of the protein and did not promote global protein unfolding. Our results strongly suggest that in β2-m the parts of the molecule involved in aggregation are located in preformed β-strands. Therefore, it is possible that local unfolding events may allow anomalous intermolecular interaction between this preformed elements leading to the formation of an aggregated β-sheet structure. This would explain the formation of amyloid deposits in hemodyalisis patients in which no major unfolding of the protein is expected to occur, as well as the effect of seeds, which may have exposed aggregation prone β-strands, in strongly accelerating the aggregation process of β2-m under physiological conditions [55].

Lysozyme

Human lysozyme has been shown to form amyloid fibrils in individuals suffering from nonneuropathic systemic amyloidosis. The disease is always associated to point mutations in the lysozyme gene and fibrils are deposited widely in tissues [56]. The properties of two amyloidogenic lysozyme mutants (I56T and D67H) have been studied in detail and, when compared to those of the wild-type protein, the mutants were found to have reduced structural stability allowing unfolding to take place at least partially at physiologically relevant temperatures [57,58]. Thus, the formation of amyloid fibrils by human lysozyme is likely to occur by the exposure of aggregation-prone region previously hidden in the native structure. The aggregation profile of lysozyme identifies three main "hot spots" corresponding to residues 20–34, 50–62 and 73–104 (Fig. 3). The last large aggregation-prone region includes several local maxima. The first "hot spot" maps in helix B, the second in a β-hairpin of the β-domain and the third includes helix C and a large flanking unstructured region at its N-terminus (Fig. 2). Although there is no experimental characterization of amyloidogenic regions in human lysozyme in the literature, this information is available for the homologous hen lysozyme molecule, which displays an almost identical 3D-structure. The aggregation profile for the hen protein is very similar to that of the human one despite the fact that our input consists solely on the sequence and the identity between both molecules is only of 40%. The equivalent "hot spots" in hen lysozyme comprise residues 24–34, 50–62 and 76–98. Experimental data suggests that the sequence of the β-domain could be of particular relevance for lysozyme aggregation since it unfolds prior to the α-domain [58]. Two peptides encompassing the β-domain of native lysozyme displayed very different behavior: peptide 61–82 appeared to be predominantly unstructured whereas peptide 41–60 showed a high tendency to aggregate and form extended β-sheet structures [59]. The first peptide coincides with a region of very low aggregation propensity in the aggregation profiles, whereas the second one covers the region with the highest aggregation propensity in the profile (residues 50–64). Interestingly enough, a peptide spanning residues 49–64 has been shown to form fibrils with the typical structure of amyloid showing that the first residues of the 41–60 peptide are not relevant for aggregation, as predicted by our approach [60]. Another study has reported that the major fragment incorporated in the core of the fibril structure, as monitored using proteolysis, encompasses the chain region 49–101 [61]. These lysozyme fragments contain helix C and two of the three β-strands of the β-domain of the native protein structure and coincide with the limits of the second and third regions in our predictions (Fig. 2 and 3). This observation could be biologically relevant, since the β-domain and C-helix of the human lysozyme have been shown to unfold locally in the amyloidogenic variant D67H, which is associated with the familial cases of systemic amyloidosis linked to lysozyme deposition [58]. The C-helix is the α-helix with the lowest helical propensity of hen lysozyme according to both theoretical and peptide based studies [59]. This low propensity might be related to the ability of this region to be incorporated into the β-sheet rich fibrillar structured as have been reported for other protein systems [62]. Limited proteolysis of hen lysozyme renders fragments 57–107 and 1–38/108–129 [61]. In the 1–38/108–129 fragment the N-terminal and C-terminal ends of the molecule are joined by a disulfide bond. Only fragment 57–107, but not fragment 1–38/108–129, is able to generate well defined amyloid [61]. Whereas the behavior of the 57–107 fragment is expected from the analysis, one should also expect the fragment 1–38 to have a high tendency to aggregate. Two explanations are possible to account for this discordance. First, it could occur that the helical structure of this region prevents its conversion to β-sheet conformation, since the A-helix displays the highest helical propensity out of all lysozyme helices [59]. The second possibility is that, being joined to the 108–129 region, predicted to have lower aggregation propensity, steric hindrances limit self-assembly or alternatively the average aggregation tendency of this peptide becomes reduced. The analysis supports this last hypothesis reporting a decrease in aggregation propensity (ΔAP = -5.34) in the joined peptide respect the 1–38 peptide alone.

Transthyretin

Transthyretin (TTR) is a homotetramer of 127-amino acid subunits. TTR is found in human plasma and cerebral spinal fluid, the plasma form being the amyloidogenetic precursor. TTR constitutes the fibrillar protein found in familial amyloidotic polyneuropathy (FAP) and senile systemic amyloidosis (SSA) [63]. In the case of FAP, the amyloid is associated with a point mutation in the TTR gene. To date, 100 different TTR mutations have been reported, many of which are amyloidogenic [64]. The FAP-associated variants characterized thus far although tetrameric, are destabilized [65]. This destabilization allows tetramer dissociation to the amyloidogenic monomeric intermediate to occur under the influence of mild denaturing denaturation conditions. More than 10 FAP-related variants crystal structures have been solved, revealing that the tertiary and quaternary structures are essentially identical to the wild type form [65]. This observation suggests that the partial denaturation of TTR is a requirement for amyloidogenesis. In this state, the presence of "hot spots" of aggregation could play an especially important role in promoting/driving amyloid formation. According to the analysis of the aggregation profile shown in Fig. 3, the TTR monomer displays three main "hot spots" encompassing residues 10–20, 23–33 and 105–118. Also in this case, aggregation-prone sequences appear to be located in preformed β-sheet structures: A β-strand (11–19), part of the B β-strand (28–36) and G and beginning of H β-strands (104–123) (Fig. 3). Most of these secondary structure elements are involved in the formation of the tetrameric structure: H strands mediate the dimerization whereas A and G provide the contacts for the tetramerization of two preformed dimmers. This explains the protective role played by the TTR quaternary structure against aggregation, since it hides or blocks most of the aggregation prone regions. Dissociation of the tetramer has been reported as a prerequisite for amyloidosis and according to our results might be associated to the exposure of previously hidden amyloidogenic sequences. We detect several short peaks exhibiting high aggregation propensities in the central region (63–94) of TTR. These result from the presence of almost regularly placed residues with low aggregation propensity (Asp, Glu, Arg, Lys, Gly) in this rather hydrophobic sequence, which probably act as disrupters, significantly lowering the aggregation tendency of this particular region, a strategy suggested to be used by nature to avoid edge-to-edge aggregation [66].

To date two different fragments of TTR have been shown to form amyloid fibrils. The peptide 105–115 can be assembled into homogeneous amyloid fibrils with favorable spectroscopic properties [67]. This has allowed to solve its fibrillar structure at high-resolution, showing that it adopts an antiparallel extended beta-strand conformation in the amyloid fibrils [68]. This peptide coincides with the region with the highest aggregation propensity in the profile. Also in excellent agreement with the prediction, the peptide 10–20 is the only other fragment of TTR reported to form amyloid fibrils [69]. No data are available on the region 23–33 but the success of the present method in predicting relevant regions in TTR suggests that it is worth to characterize its *in vitro* aggregation capabilities.

Prion protein

Misfolded isoforms of the naturally occurring prion protein (PrP) have been shown to be the causative agents in many mammalian neurodegenerative disorders, including Cruetzfeldt-Jakob disease (CJD) in human, scrapie in sheep, and bovine spongiform encephalopathy in cows. Prion infectivity is unique in that the pathogenic prion form (PrP^{Sc}) is involved in the conversion of the endogenous conformation (PrP^{C}) into transformed PrP^{Sc}. The "protein-only" hypothesis [70] asserts further that no extraneous agents are necessary to explain the unusual behavior of prions. Prion diseases can have infectious, familial, and sporadic origins. The basic infectious mechanism is thought to be a conformational change of the normal prion protein (PrPC) into the pathogenic PrP^{Sc} catalyzed by PrP^{Sc} itself.

The normal prion protein (PrP^{C}) is a GPI-anchored glycoprotein constitutively expressed on the surface of primarily neuronal cells. It consists of two structurally different parts; a C-terminal, globular part mainly α-helical in nature (Fig. 2) and an unstructured, N-terminal part [71]. Misfolding of PrP^{C} into PrP^{Sc} occurs posttranslationally and results in increased β-sheet content and gain of protease-resistance. Fig. 3 shows the predicted "hot spots" in the aggregation profile of the full-length human prion protein. They are located at the N-terminus (1–32), in the central region (105–146) and the C-terminus (208–252), respectively.

The role of the detected aggregation-prone sequence at the N-terminus is uncertain since it is out of the protease resistant core of PrP^{Sc}. Little information exits about the role of this region, although it appears to be unnecessary both for prion transmission and aggregation. The predicted C-terminal "hot spot" includes almost all the C-terminal α-helix, named C, from the globular domain (Fig. 2). Interestingly, some of the human mutations linked to Creutzfeldt-Jakob disease occur in this region of the prion protein and it has been related to the conversion of PrP^{C} into the toxic PrP^{Sc}. Moreover, some strains of PrP resistant to conversion to PrP^{Sc} have been found to bear mutations in helix C, and positions 214 and 218 have been shown to modulate PrP^{Sc} formation [72]. It is also important to note that the main structural differences between prion proteins from different species have been found at the end of helix C [71].

The central region of PrPC linking the unstructured N-terminal part with the globular C-terminal domain is believed to play a pivotal role in the PrP^{C} conformational changes. Extensive studies on the secondary structure and fibrillogenic properties of synthetic peptides of PrP have established that the continuous segment of the prion protein spanning residues 106–147, coincident with the second "hot spot" predicted using our approach, is important for the fibrillogenic properties of the protein [73]. One of the synthetic peptides, that named PrP106-126 within the central region of PrP and near the N-terminal of the protease resistant core of PrP^{Sc}, shares many properties with the infectious form as it readily forms amyloid fibrils with a high β-sheet content, shows partial proteinase K resistance and is neurotoxic *in vivo* [74]. The neurotoxicity of PrP106-126 depends on the expression of endogenous PrP^{C} which makes PrP106-126 a relevant model for PrP^{Sc} neurotoxicity [74]. Also another prion derived peptide – PrP118-135 – has been found to cause neuronal death via induction of apoptosis [75]. The toxicity of PrP118-135 is, however, independent of endogenous PrP^{C} expression. Both peptides map in our predicted central aggregation-prone region of PrP^{C}.

Conclusion

Overall, the method described here appears as a useful tool for the identification of protein regions that are especially relevant for protein aggregation and amyloidogenesis both in natively unfolded and properly folded globular proteins (Table 4). The results provide support to the hypothesis that short specific amino acid stretches can act as triggers for the incorporation of polypeptides into amyloid structures. It is interesting to note that in those cases in which structural information allows to delimitate the region incorporated in the core of the fibrillar structure, our predicted "hot spots" and those proved experimentally are considerably shorter than the whole region, suggesting that the role of "hot spots" is to act as specific nucleation points from which the ordered fibrillar structure is expanded.

Nature has provided globular proteins with a reasonable conformational stability in the native state in which, as proved here, aggregation-prone sequences are buried or involved in intra-molecular interactions. This appears as a very successful evolutive strategy to avoid aggregation, since few proteins aggregate from their stable native conformation. Accordingly, amyloid-related mutations in globular proteins usually result in destabilization of the folded state allowing the exposure of previously hidden "hot spots", as those reported here. This explains the scarce success in predicting the effect of mutations in the aggregation of globular proteins (data not shown), whereas the prediction of fatal sequence changes in intrinsically unstructured proteins involved in disease is gener-

Table 4: List of the predicted "hot spots" in the different disease-linked polypeptides in this study and comparison with the available experimental data. Experimental "hot spots" refer to those protein regions shown to be involved in the aggregation process of the corresponding polypeptide. It is also noted if the predicted "hot spot" has been described as a structural element of the amyloid fibrils formed by the different peptides and proteins in the study.

Protein	Predicted "Hot Spots"	Experimental "Hot Spots"	Regions in the fibrils
Amyloid-β-protein	16–21	+	+
	30–36	+	+
	38–42	+	+
Islet amyloid polypeptide	12–18	+	uncertain
	22–28	+	uncertain
	1–18	No experimental data available	uncertain
α-Synuclein	27–56	+	uncertain
	61–94	+	+
β2-Microgobulin	21–31	+	+
	56–69	+	+
	79–85	+	+
	87–91	+	+
Lysozyme (hen)	24–34	-	-
	50–62	+	+
	76–98	+	+
Transthyretin	10–20	+	+
	23–33	No experimetal data available	uncertain
	105–118	+	+
Prion Protein	1–32	No experimetal data available	uncertain
	105–146	+	+
	208–252	No experimetal data available	uncertain

ally accurate. The effects of such mutations can be explained in most cases by intrinsic factors, as they directly result in changes on the average propensity of the full polypeptide to aggregate.

Besides providing important clues about the mechanism of protein aggregation, this study may be relevant for the therapeutics of amyloid disease, since the identified "hot spots" could be regarded as preferential targets to tackle the deleterious disorders linked to protein deposition. According to our results, different specific strategies should be employed when designing methods to avoid aggregation, depending on the disease being caused by natively unfolded or by globular proteins. In Alzheimer, type II diabetes and Parkinson diseases, shielding the already exposed aggregation-prone regions in the polypeptides by using small compounds or antibodies appears as a promising approach, whereas compounds that will stabilize the native conformation and avoid the exposure of the deleterious "hot spots" will be more effective in the case of globular proteins. Additionally, when gene therapy eventually comes to age, mutations that disrupt aggregation-prone regions in unstructured polypeptides or those which over-stabilize the native state of globular aggregation-prone proteins are expected to be useful approaches to avoid protein deposition and meliorate neurodegenerative and systemic amyloidogenic disorders.

Methods

Experimental determination of amino acids aggregation propensities

The CHC of Aβ42 peptide was chosen as a paradigmatic aggregation-prone region for the calculation of the individual effect of each natural amino acid on protein aggregation. The specific effect on Aβ42's deposition promoted by the 20 different natural amino acids when located in the central position of this model "hot spot" were evaluated. Briefly, the wild type Aβ42 gene and its 19 mutants were inserted as a fusion protein upstream of the green fluorescence protein (GFP) and expressed individually in bacteria. In this system, the levels of GFP fluorescence in the cells depend exclusively on the *in vivo* aggregation propensity of the Aβ42 variant [10,25], in such a way that changes in aggregation propensities promoted by the different mutations can be easily monitored by measuring the fluorescence emission of the cells expressing each particular variant and normalizing it relative to that emitted by the cells bearing the wild type sequence. Three independent clones were analyzed for each mutation and each clone was analyzed at least by triplicate to generate consistent data. To obtain the individual aggregation propensities in Table 1, the change promoted by each amino acid was normalized relative to the average change of the pool of 20 amino acids.

Generation of aggregation profiles and identification of "hot spots"

Different experimental data suggest that the aggregation of Aβ42 occurs from a mostly unfolded conformation in which the CHC is exposed to solvent [76]. Assuming that the individual intrinsic aggregation propensities obtained analyzing this particular protein region will probably apply for any unfolded sequence; an aggregation profile was generated for every protein in this study through a simple assignment of the values in Table 1 to each individual residue in the corresponding sequences. Since "hot spots" are clusters of consecutive residues, the sequence was scanned by using a five residues sliding window. "Hot spots" in the sequence were identified as those protein regions at least five residues in length (the minimal size shown to date to be required for a peptide to form amyloid fibrils similar to those formed by whole polypeptides [77], in which the aggregation propensity is above the average aggregation propensity of the complete sequence. The average propensity of the polypeptide was calculated as the sum of the aggregation propensities of its individual amino acids divided by the number of residues.

Analysis of the effect of changes in the polypeptide sequence on aggregation

The concept of "hot spot" of aggregation implies that the contribution of a particular residue in a protein sequence on protein aggregation is somehow modulated by its immediate neighbors. According to this, the effects of mutation on protein aggregation can not be properly calculated by a simple subtraction of the intrinsic aggregation propensities of the wild type and mutant residues. Instead, to provide a more general description of the effect of the change on the overall aggregation propensity, the individual aggregation profiles for the wild type protein and the different mutants are obtained and the differences between the areas below the corresponding profiles are calculated. The area between each profile was always normalized by the number of residues in the considered species to compare between the aggregation propensities of the complete protein and fragments coming from proteolysis, chemical synthesis or other processes. The difference between normalized areas, multiplied by a 100 factor, was designed as the change in average aggregation propensity (ΔAP). ΔAP will be positive if the mutation is predicted to increase the aggregation propensity of the polypeptide chain and negative if it is predicted to increase solubility.

Abbreviations

AD Alzheimer's disease

Aβ Amyloid-β-protein

CHC Central hydrophobic cluster

FAD Familial Alzheimer diseases

FAP Amyloidotic polyneuropathy

GFP Green fluorescent protein

IAPP Islet amyloid polypeptide

NAC Non-Aβ component of amyloid plaques

PrP Prion protein

PrP^{Sc} Pathogenic prion form

SSA Senile systemic amyloidosis

TTR Transthyretin

ΔAP Change in average aggregation propensity

β2-m β2-Microglobulin

Authors' contributions

NSG and IP performed most of the experiments and prepared the final data and figures. FXA and JV contributed to data interpretation and manuscript redaction. SV directed the work and prepared the manuscript.

Acknowledgements

The vector expressing the Aβ42-GFP fusion was a generous gift of Michael Hecht's group. This work has been supported by Grants BIO2001-2046 and BIO2004-05879 (Ministerio de Ciencia y Tecnología, MCYT, Spain), by the Centre de Referència en Biotecnologia (Generalitat de Catalunya, Spain) and by PNL2004-40 (Universitat Autònoma de Barcelona (UAB). S.V. is supported by a "Ramón y Cajal" project awarded by the MCYT and co-financed by the UAB.

References

1. Smith A: **protein misfolding.** *Nature* 2003, **426:**883-883.
2. Rochet JC, Lansbury PT Jr: **Amyloid fibrillogenesis: themes and variations.** *Curr Opin Struct Biol* 2000, **10:**60-68.
3. Uversky VN, Fink AL: **Conformational constraints for amyloid fibrillation: the importance of being unfolded.** *Biochim Biophys Acta* 2004, **1698:**131-153.
4. Kelly JW: **The alternative conformations of amyloidogenic proteins and their multi-step assembly pathways.** *Curr Opin Struct Biol* 1998, **8:**101-106.
5. Ventura S, Zurdo J, Narayanan S, Parreno M, Mangues R, Reif B, Chiti F, Giannoni E, Dobson CM, Aviles FX, Serrano L: **Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case.** *Proc Natl Acad Sci U S A* 2004, **101:**7258-7263.
6. Ivanova MI, Sawaya MR, Gingery M, Attinger A, Eisenberg D: **An amyloid-forming segment of beta2-microglobulin suggests a molecular model for the fibril.** *Proc Natl Acad Sci U S A* 2004, **101:**10584-10589.
7. Dobson CM: **Protein misfolding, evolution and disease.** *Trends Biochem Sci* 1999, **24:**329-332.
8. Linding R, Schymkowitz J, Rousseau F, Diella F, Serrano L: **A comparative study of the relationship between protein structure

and beta-aggregation in globular and intrinsically disordered proteins. *J Mol Biol* 2004, **342:**345-353.

9. Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L: **Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins.** *Nat Biotechnol* 2004, **22:**1302-1306.
10. de Groot N, Aviles FX, Vendrell J, Ventura S: 2005. Submitted
11. Bitan G, Vollers SS, Teplow DB: **Elucidation of primary structure elements controlling early amyloid beta-protein oligomerization.** *J Biol Chem* 2003, **278:**34882-34889.
12. Chiti F, Stefani M, Taddei N, Ramponi G, Dobson CM: **Rationalization of the effects of mutations on peptide and protein aggregation rates.** *Nature* 2003, **424:**805-808.
13. Fink AL: **Protein aggregation: folding aggregates, inclusion bodies and amyloid.** *Fold Des* 1998, **3:**R9-23.
14. Pawar AP, Dubay KF, Zurdo J, Chiti F, Vendruscolo M, Dobson CM: **Prediction of "Aggregation-prone" and "Aggregation-susceptible" Regions in Proteins Associated with Neurodegenerative Diseases.** *J Mol Biol* 2005, **350:**379-392.
15. Selkoe DJ: **Alzheimer's disease: genes, proteins, and therapy.** *Physiol Rev* 2001, **81:**741-766.
16. Nagele RG, Wegiel J, Venkataraman V, Imaki H, Wang KC, Wegiel J: **Contribution of glial cells to the development of amyloid plaques in Alzheimer's disease.** *Neurobiol Aging* 2004, **25:**663-674.
17. Williams AD, Portelius E, Kheterpal I, Guo JT, Cook KD, Xu Y, Wetzel R: **Mapping abeta amyloid fibril secondary structure using scanning proline mutagenesis.** *J Mol Biol* 2004, **335:**833-842.
18. Morimoto A, Irie K, Murakami K, Masuda Y, Ohigashi H, Nagao M, Fukuda H, Shimizu T, Shirasawa T: **Analysis of the secondary structure of beta-amyloid (Abeta42) fibrils by systematic proline replacement.** *J Biol Chem* 2004, **279:**52781-52788.
19. Petkova AT, Ishii Y, Balbach JJ, Antzutkin ON, Leapman RD, Delaglio F, Tycko R: **A structural model for Alzheimer's beta-amyloid fibrils based on experimental constraints from solid state NMR.** *Proc Natl Acad Sci U S A* 2002, **99:**16742-16747.
20. Torok M, Milton S, Kayed R, Wu P, McIntire T, Glabe CG, Langen R: **Structural and dynamic features of Alzheimer's Abeta peptide in amyloid fibrils studied by site-directed spin labeling.** *J Biol Chem* 2002, **277:**40810-40815.
21. Balbach JJ, Ishii Y, Antzutkin ON, Leapman RD, Rizzo NW, Dyda F, Reed J, Tycko R: **Amyloid fibril formation by A beta 16–22, a seven-residue fragment of the Alzheimer's beta-amyloid peptide, and structural characterization by solid state NMR.** *Biochemistry* 2000, **39:**13748-13759.
22. Findeis MA, Musso GM, Arico-Muendel CC, Benjamin HW, Hundal AM, Lee JJ, Chin J, Kelley M, Wakefield J, Hayward NJ, Molineaux SM: **Modified-peptide inhibitors of amyloid beta-peptide polymerization.** *Biochemistry* 1999, **38:**6791-6800.
23. Jarrett JT, Berger EP, Lansbury PT Jr: **The carboxy terminus of the beta amyloid protein is critical for the seeding of amyloid formation: implications for the pathogenesis of Alzheimer's disease.** *Biochemistry* 1993, **32:**4693-4697.
24. Yamamoto N, Hasegawa K, Matsuzaki K, Naiki H, Yanagisawa K: **Environment- and mutation-dependent aggregation behavior of Alzheimer amyloid beta-protein.** *J Neurochem* 2004, **90:**62-9.
25. Wurth C, Guimard NK, Hecht MH: **Mutations that reduce aggregation of the Alzheimer's Abeta42 peptide: an unbiased search for the sequence determinants of Abeta amyloidogenesis.** *J Mol Biol* 2002, **319:**1279-1290.
26. Clark A, Cooper GJ, Lewis CE, Morris JF, Willis AC, Reid KB, Turner RC: **Islet amyloid formed from diabetes-associated peptide may be pathogenic in type-2 diabetes.** *Lancet* 1987, **2:**231-234.
27. Goldsbury C, Goldie K, Pellaud J, Seelig J, Frey P, Muller SA, Kistler J, Cooper GJ, Aebi U: **Amyloid fibril formation from full-length and fragments of amylin.** *J Struct Biol* 2000, **130:**352-362.
28. Kajava AV, Aebi U, Steven AC: **The parallel superpleated beta-structure as a model for amyloid fibrils of human amylin.** *J Mol Biol* 2005, **348:**247-252.
29. Scrocchi LA, Ha K, Chen Y, Wu L, Wang F, Fraser PE: **Identification of minimal peptide sequences in the (8–20) domain of human islet amyloid polypeptide involved in fibrillogenesis.** *J Struct Biol* 2003, **141:**218-27.
30. Tracz SM, Abedini A, Driscoll M, Raleigh DP: **Role of aromatic interactions in amyloid formation by peptides derived from human Amylin.** *Biochemistry* 2004, **43:**15901-15908.
31. Moriarty DF, Raleigh DP: **Effects of sequential proline substitutions on amyloid formation by human amylin 20-29.** *Biochemistry* 1999, **38:**1811-1818.
32. Azriel R, Gazit E: **Analysis of the minimal amyloid-forming fragment of the islet amyloid polypeptide. An experimental support for the key role of the phenylalanine residue in amyloid formation.** *J Biol Chem* 2001, **276:**34156-34161.
33. Green J, Goldsbury C, Mini T, Sunderji S, Frey P, Kistler J, Cooper G, Aebi U: **Full-length rat amylin forms fibrils following substitution of single residues from human amylin.** *J Mol Biol* 2003, **326:**1147-1156.
34. Sakagashira S, Sanke T, Hanabusa T, Shimomura H, Ohagi S, Kumagaye KY, Nakajima K, Nanjo K: **Missense mutation of amylin gene (S20G) in Japanese NIDDM patients.** *Diabetes* 1996, **45:**1279-1281.
35. Porte D Jr, Kahn SE: **Hyperproinsulinemia and amyloid in NIDDM. Clues to etiology of islet beta-cell dysfunction?** *Diabetes* 1989, **38:**1333-1336.
36. Spillantini MG, Schmidt ML, Lee VM, Trojanowski JQ, Jakes R, Goedert M: **Alpha-synuclein in Lewy bodies.** *Nature* 1997, **388:**839-840.
37. Goedert M: **Alpha-synuclein and neurodegenerative diseases.** *Nat Rev Neurosci* 2001, **2:**492-501.
38. Bodles AM, Guthrie DJ, Greer B, Irvine GB: **Identification of the region of non-Abeta component (NAC) of Alzheimer's disease amyloid responsible for its aggregation and toxicity.** *J Neurochem* 2001, **78:**384-395.
39. Giasson BI, Murray IV, Trojanowski JQ, Lee VM: **A hydrophobic stretch of 12 amino acid residues in the middle of alpha-synuclein is essential for filament assembly.** *J Biol Chem* 2001, **276:**2380-2386.
40. Miake H, Mizusawa H, Iwatsubo T, Hasegawa M: **Biochemical characterization of the core structure of alpha-synuclein filaments.** *J Biol Chem* 2002, **277:**19213-10219.
41. Der-Sarkissian A, Jao CC, Chen J, Langen R: **Structural organization of alpha-synuclein fibrils studied by site-directed spin labeling.** *J Biol Chem* 2003, **278:**37530-37535.
42. Choi W, Zibaee S, Jakes R, Serpell LC, Davletov B, Crowther RA, Goedert M: **Mutation E46K increases phospholipid binding and assembly into filaments of human alpha-synuclein.** *FEBS Lett* 2004, **576:**363-8.
43. Cuervo AM, Stefanis L, Fredenburg R, Lansbury PT, Sulzer D: **Impaired degradation of mutant alpha-synuclein by chaperone-mediated autophagy.** *Science* 2004, **305:**1292-1295.
44. Saha AR, Hill J, Utton MA, Asuni AA, Ackerley S, Grierson AJ, Miller CC, Davies AM, Buchman VL, Anderton BH, Hanger DP: **Parkinson's disease alpha-synuclein mutations exhibit defective axonal transport in cultured neurons.** *J Cell Sci* 2004, **117:**1017-1024.
45. Huang Y, Cheung L, Rowe D, Halliday G: **Genetic contributions to Parkinson's disease.** *Brain Res Brain Res Rev* 2004, **46:**44-70.
46. Koch KM: **Dialysis-related amyloidosis.** *Kidney Int* 1992, **41:**1416-1429.
47. McParland VJ, Kad NM, Kalverda AP, Brown A, Kirwin-Jones P, Hunter MG, Sunde M, Radford SE: **Partially unfolded states of beta(2)-microglobulin and amyloid formation in vitro.** *Biochemistry* 2000, **39:**8735-8746.
48. Kozhukh GV, Hagihara Y, Kawakami T, Hasegawa K, Naiki H, Goto Y: **Investigation of a peptide responsible for amyloid fibril formation of beta 2-microglobulin by achromobacter protease I.** *J Biol Chem* 2002, **277:**1310-1315.
49. Hasegawa K, Ohhashi Y, Yamaguchi I, Takahashi N, Tsutsumi S, Goto Y, Gejyo F, Naiki H: **Amyloidogenic synthetic peptides of beta2-microglobulin–a role of the disulfide bond.** *Biochem Biophys Res Commun* 2003, **304:**101-106.
50. Jones S, Manning J, Kad NM, Radford SE: **Amyloid-forming peptides from beta2-microglobulin-Insights into the mechanism of fibril formation in vitro.** *J Mol Biol* 2003, **325:**249-57.
51. Ivanova MI, Gingery M, Whitson LJ, Eisenberg D: **Role of the C-terminal 28 residues of beta2-microglobulin in amyloid fibril formation.** *Biochemistry* 2003, **42:**13536-13540.
52. Bellotti V, Stoppini M, Mangione P, Sunde M, Robinson C, Asti L, Brancaccio D, Ferri G: **Beta2-microglobulin can be refolded into a**

native state from ex vivo amyloid fibrils. *Eur J Biochem* 1998, **258:**61-67.

53. Ivanova MI, Sawaya MR, Gingery M, Attinger A, Eisenberg D: **An amyloid-forming segment of beta2-microglobulin suggests a molecular model for the fibril.** *Proc Natl Acad Sci U S A* 2004, **101:**10584-10589.
54. Yamamoto S, Hasegawa K, Yamaguchi I, Tsutsumi S, Kardos J, Goto Y, Gejyo F, Naiki H: **Low concentrations of sodium dodecyl sulfate induce the extension of beta 2-microglobulin-related amyloid fibrils at a neutral pH.** *Biochemistry* 2004, **43:**11075-11082.
55. Kihara M, Chatani E, Sakai M, Hasegawa K, Naiki H, Goto Y: **Seeding-dependent maturation of beta2-microglobulin amyloid fibrils at neutral pH.** *J Biol Chem* 2005, **280:**12012-12018.
56. Pepys MB, Hawkins PN, Booth DR, Vigushin DM, Tennent GA, Soutar AK, Totty N, Nguyen O, Blake CCF, Terry CJ, Feest TG, Zalin AM, Hsuan JJ: **Human lysozyme gene mutations cause hereditary systemic amyloidosis.** *Nature* 1993, **362:**553-557.
57. Booth DR, Sunde M, Bellotti V, Robinson CV, Hutchinson WL, Fraser PE, Hawkins PN, Dobson CM, Radford SE, Blake CC, Pepys MB: **Instability, unfolding and aggregation of human lysozyme variants underlying amyloid fibrillogenesis.** *Nature* 1997, **385:**787-793.
58. Canet D, Last AM, Tito P, Sunde M, Spencer A, Archer DB, Redfield C, Robinson CV, Dobson CM: **Local cooperativity in the unfolding of an amyloidogenic variant of human lysozyme.** *Nat Struct Biol* 2002, **9:**308-315.
59. Yang JJ, Buck M, Pitkeathly M, Kotik M, Haynie DT, Dobson CM, Radford SE: **Conformational properties of four peptides spanning the sequence of hen lysozyme.** *J Mol Biol* 1995, **29:**483-491.
60. Krebs MR, Wilkins DK, Chung EW, Pitkeathly MC, Chamberlain AK, Zurdo J, Robinson CV, Dobson CM: **Formation and seeding of amyloid fibrils from wild-type hen lysozyme and a peptide fragment from the beta-domain.** *J Mol Biol* 2000, **300:**541-549.
61. Frare E, Polverino De Laureto P, Zurdo J, Dobson CM, Fontana A: **A highly amyloidogenic region of hen lysozyme.** *J Mol Biol* 2004, **340:**1153-1165.
62. Kallberg Y, Gustafsson M, Persson B, Thyberg J, Johansson J: **Prediction of amyloid fibril-forming proteins.** *J Biol Chem* 2001, **276:**12945-12950.
63. Saraiva MJ, Birken S, Costa PP, Goodman DS: **Amyloid fibril protein in familial amyloidotic polyneuropathy, Portuguese type. Definition of molecular abnormality in transthyretin (prealbumin).** *J Clin Invest* 1984, **74:**104-119.
64. McCutchen SL, Lai Z, Miroy GJ, Kelly JW, Colon W: **Comparison of lethal and nonlethal transthyretin variants and their relationship to amyloid disease.** *Biochemistry* 1995, **34:**13527-13536.
65. Hornberg A, Eneqvist T, Olofsson A, Lundgren E, Sauer-Eriksson AE: **A comparative analysis of 23 structures of the amyloidogenic protein transthyretin.** *J Mol Biol* 2000, **22:**649-669.
66. Richardson JS, Richardson DC: **Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation.** *Proc Natl Acad Sci U S A* 2002, **99:**2754-2759.
67. Jaroniec CP, MacPhee CE, Astrof NS, Dobson CM, Griffin RG: **Molecular conformation of a peptide fragment of transthyretin in an amyloid fibril.** *Proc Natl Acad Sci U S A* 2002, **99:**16748-16753.
68. Jaroniec CP, MacPhee CE, Bajaj VS, McMahon MT, Dobson CM, Griffin RG: **High-resolution molecular structure of a peptide in an amyloid fibril determined by magic angle spinning NMR spectroscopy.** *Proc Natl Acad Sci U S A* 2004, **101:**711-716.
69. Jarvis JA, Kirkpatrick A, Craik DJ: **1H NMR analysis of fibril-forming peptide fragments of transthyretin.** *Int J Pept Protein Res* 1994, **44:**388-39.
70. Prusiner SB, Scott MR, DeArmond SJ, Cohen FE: **Prion protein biology.** *Cell* 1998, **93:**337-348.
71. Zahn R, Liu A, Luhrs T, Riek R, von Schroetter C, Lopez Garcia F, Billeter M, Calzolai L, Wider G, Wuthrich K: **NMR solution structure of the human prion protein.** *Proc Natl Acad Sci U S A* 2000, **97:**145-150.
72. Kaneko K, Zulianello L, Scott M, Cooper CM, Wallace AC, James TL, Cohen FE, Prusiner SB: **Evidence for protein X binding to a discontinuous epitope on the cellular prion protein during scrapie prion propagation.** *Proc Natl Acad Sci U S A* 1997, **94:**10069-10074.
73. Tagliavini F, Prelli F, Verga L, Giaccone G, Sarma R, Gorevic P, Ghetti B, Passerini F, Ghibaudi E, Forloni G, Salmona M, Bugiani O, Frangione B: **Synthetic peptides homologous to prion protein residues 106–147 form amyloid-like fibrils in vitro.** *Proc Natl Acad Sci U S A* 1993, **90:**9678-9682.
74. Gu Y, Fujioka H, Mishra RS, Li R, Singh N: **Prion peptide 106–126 modulates the aggregation of cellular prion protein and induces the synthesis of potentially neurotoxic transmembrane PrP.** *J Biol Chem* 2002, **277:**2275-86.
75. Chabry J, Ratsimanohatra C, Sponne I, Elena PP, Vincent JP, Pillot T: **In vivo and in vitro neurotoxicity of the human prion protein (PrP) fragment P118-135 independently of PrP expression.** *J Neurosci* 2003, **23:**462-469.
76. Santini S, Wei G, Mousseau N, Derreumaux P: **Pathway complexity of Alzheimer's beta-amyloid Abeta16-22 peptide assembly.** *Structure* 2004, **12:**1245-1255.
77. Haspel N, Zanuy D, Ma B, Wolfson H, Nussinov R: **A comparative study of amyloid fibril formation by residues 15–19 of the human calcitonin hormone: a single beta-sheet model with a small hydrophobic core.** *J Mol Biol* 2005, **345:**1213-1227.

BMC Bioinformatics

Software

AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides

Oscar Conchillo-Solé†[1], Natalia S de Groot†[2], Francesc X Avilés[1,2], Josep Vendrell[1,2], Xavier Daura[1,3] and Salvador Ventura*[1,2]

Address: [1]Institut de Biotecnologia i de Biomedicina (IBB), Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain, [2]Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain and [3]Catalan Institution for Research and Advanced Studies (ICREA), 08010 Barcelona, Spain

Email: Oscar Conchillo-Solé - Oscar.Conhillo@bioinf.uab.cat; Natalia S de Groot - natalia.sanchez@uab.es; Francesc X Avilés - francescxavier.aviles@uab.es; Josep Vendrell - josep.vendrell@uab.es; Xavier Daura - xavier.daura@uab.es; Salvador Ventura* - salvador.ventura@uab.es

* Corresponding author †Equal contributors

Published: 27 February 2007

BMC Bioinformatics 2007, **8**:65 doi:10.1186/1471-2105-8-65

Received: 22 November 2006
Accepted: 27 February 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/65

Abstract

Background: Protein aggregation correlates with the development of several debilitating human disorders of growing incidence, such as Alzheimer's and Parkinson's diseases. On the biotechnological side, protein production is often hampered by the accumulation of recombinant proteins into aggregates. Thus, the development of methods to anticipate the aggregation properties of polypeptides is receiving increasing attention. AGGRESCAN is a web-based software for the prediction of aggregation-prone segments in protein sequences, the analysis of the effect of mutations on protein aggregation propensities and the comparison of the aggregation properties of different proteins or protein sets.

Results: AGGRESCAN is based on an aggregation-propensity scale for natural amino acids derived from *in vivo* experiments and on the assumption that short and specific sequence stretches modulate protein aggregation. The algorithm is shown to identify a series of protein fragments involved in the aggregation of disease-related proteins and to predict the effect of genetic mutations on their deposition propensities. It also provides new insights into the differential aggregation properties displayed by globular proteins, natively unfolded polypeptides, amyloidogenic proteins and proteins found in bacterial inclusion bodies.

Conclusion: By identifying aggregation-prone segments in proteins, AGGRESCAN http://bioinf.uab.es/aggrescan/ shall facilitate (*i*) the identification of possible therapeutic targets for anti-depositional strategies in conformational diseases and (*ii*) the anticipation of aggregation phenomena during storage or recombinant production of bioactive polypeptides or polypeptide sets.

Background

Protein aggregation has become a key topic in both biotechnological and medical sciences [1,2]. It constitutes the main bottleneck in protein production, narrowing the spectrum of relevant polypeptides obtained by recombinant techniques [3]; it reduces the shelf life and increases the immunogenicity of polypeptidic drugs [4]; and it is associated with an increasing number of critical human diseases including Alzheimer's disease, spongiform encephalopaties, type II diabetes mellitus and Parkinson's disease [5-8].

In the last decade data have begun to accumulate suggesting that the composition and the primary structure of a polypeptide determine to a large extent its propensity to aggregate and that small changes may have a huge impact on solubility. The ability to predict the aggregation propensity of a protein from its sequence would be of much value, for example, in the control of unwanted protein deposition events through specific sequence targeted therapeutics or in the discovery of more soluble variants of proteins of biotechnological interest. It is commonly assumed that not all regions of a polypeptide are equally important in determining its aggregation tendency. In this context, some authors have recently proved that very short specific amino acid stretches can act as facilitators or inhibitors of amyloid fibril formation [9,10]. These relevant regions are usually known as aggregation "hot spots" (HS) and their presence has been described in most of the peptides and proteins underlying neurodegenerative and systemic amyloidogenic disorders [11].

In previous work we exploited the experimental data obtained from a system *in vivo* that uses the β-amyloid peptide as model to derive a simple approach for the detection of "hot spots" of aggregation [12,13]. This approach permitted the identification of aggregation-prone segments in several unstructured and globular disease-linked polypeptides and the prediction of the effect of disease-linked mutations in some of these polypeptides. Here, we describe a software and web interface (AGGRESCAN) that implement this approach and extend it to the general prediction of aggregation "hot spots" and the evaluation of their contribution to the differential aggregation behaviour of polypeptides. In addition to enabling the simultaneous analysis of a large number of sequences, AGGRESCAN introduces a new set of functions and descriptors for the identification of "hot spots" of aggregation and the determination of their relevance within the parent sequence.

Implementation

Approach

Recent findings in the study of protein aggregation indicate that not all the polypeptides share the same aggregation propensities and that there exists specific continuous protein segments that can nucleate the aggregation process when exposed to solvent [9,10], suggesting a sequence-dependence of aggregation propensities. At the same time, it has been shown that the same physicochemical principles underlie the aggregation propensities of different polypeptides from unfolded states [14]. According to these assumptions one may expect that the conclusions obtained from the study of a relevant nucleating sequence, or "hot spot" of aggregation, in its natural polypeptidic context could apply to other unrelated proteins. Using an in vivo reporter method to study a "hot spot" in the central hydrophobic core of Aβ we calculated the effect of single point mutations on the aggregation propensities of the peptide within the cell. The results were used to approximate the *in vivo* intrinsic aggregation propensities of natural amino acids when located in an aggregation-prone sequence stretch [12] (see additional file 1). This information was subsequently used to generate an aggregation profile for any protein sequence under study to detect those regions with high aggregation propensities. Comparison of the theoretically calculated changes in aggregation propensities between a wild type sequence and different mutants serves also as a tool to predict the behavior of the mutant forms. Albeit the basic simplicity of this phenomenological model, it predicts, at least qualitatively, both the presence of experimentally validated "hot spots" and the variations in aggregation propensity introduced by mutations in some disease-related polypeptides [13].

System description

AGGRESCAN is a web-based tool with a computing core coded in C and a front end written in a combination of html and perl cgi. Development of AGGRESCAN was carried out under Mandriva Linux LE2005 and the service is currently running under Mandrake Linux 9.0 on a Pentium 4 1300 MHz (willamette) with 1GB RDRAM.

For each polypeptide sequence input, AGGRESCAN calculates and reports: *i*) an aggregation-propensity value for each residue in the sequence and a graphical representation of the profile for the entire polypeptide; *ii*) the areas of profile peaks over a precalculated threshold and a graphical representation of peak-area values; *iii*) putative aggregation "hot spots", identified from the polypeptide's aggregation profile.

Input

The polypeptide sequence(s) can be typed or pasted on screen using FASTA format. Despite supporting up to 100 characters for name entries, use of very long names is discouraged as it disturbs the visualization of the output. Sequence entries may not contain more than 2,000 residues and the letters must correspond to those associated to the 20 natural amino acids. If these two conditions are not satisfied an error message will appear on screen. White-space, enter and tab characters are ignored. Characters may be entered as lower and/or upper case, and so will remain in the output.

Processing

The calculations are based on aggregation-propensity values per amino acid (aaAV, or a3v) derived previously from experimental data [12]. The program calculates the a3v average (a4v) over a sliding window of a given length and assigns it to the central residue in the window. The size of the sliding window ([5,7,9], and [11] residues) was trained against a database of 57 amyloidogenic proteins, in which the location of "hot spots" was experimentally known. In general, the predictions of the overall aggregation-prone regions do not depend on the length of the used windows and only slightly affect their limits. There are, however, two remarkable exceptions: 1) The use of long windows on top of very short sequences results on excessive smoothing of the profile and experimentally different "hot spots" become grouped and masked and cannot be individualized in the prediction. 2) The use of short windows on top of very long sequences results in the appearance of a number of short experimentally non-relevant predicted "hot spots" with associated low areas. Thus, the procedure incorporates a ponderation of the window length relative to the size of the analyzed protein. The best predictions were obtained using a window size of 5 for < = 75 residues, 7 for < = 175, 9 for < = 300 and 11 for > 300, respectively, probably reflecting that for longer sequences larger "hot spots" are necessary in order to significantly increase their aggregation propensities, while short-stretches suffice for smaller peptides. To account for charge effects at the polypeptide's termini ($NH_3{}^+$ and COO^-) a virtual residue is added to each side of the chain (residue 0 at the N-terminus and residue n+1 at the C-terminus, n being the original sequence length). The a^3v of residue 0 equals the average a^3v of the basic residues (K, R), while that of residue n+1 equals the average a^3v of the acidic residues (D, E). The first window, ranging from residue 0 to residue 4, 6, 8 or 10 (depending on window size), will serve to assign an a^4v to residue 2, 3, 4 or 5, respectively. Thus, the off-centre residues 1, 1–2, 1–3 or 1–4 may not have an associated a^4v. This is solved by giving these residues the value corresponding to the first window centre. The same procedure is followed at the C-terminus. The "hot spot" threshold (HST) has been defined as the average of the a^3v of the 20 amino acids weighted by their frequencies in the SwissProt database [15]. The aggregation profile (AP) of the polypeptide is defined by the complete sequence of a^4v. The sum of a^4v and the average of a^3v over the entire sequence (a^4vSS and a^3vSA, respectively) are also calculated. A region in the polypeptide sequence is considered an aggregation "hot spot" (HS) if there are 5 or more sequentially continuous residues with an a^4v larger than the HST and none of them is a proline (aggregation breaker) [16]. The average a^4v in each "hot spot" is then calculated (a^4vAHS). Finally, the area of the AP above the HST (AAT), the total area (TA, HST being the zero axis), and the area above the HST of each profile peak identified as "hot spot" (HSA) are integrated numerically using the trapezoidal rule (see additional file 2).

Output

With current service resources, the delay time between pressing the submit button and receiving the output on screen is of 10 minutes for an input set of 100 sequences of sizes between 40 and 1,000 residues. The output is structured in tables, one per sequence and an additional one with averages over all sequences, an excel-readable document with output values and a list of sequences sorted by normalized a^4vSS for 100 residues (Na^4vSS). The first row in the output contains the sequence names. The second row displays links to the three graphics produced per sequence, i.e., Profile graphic: AP (red), a^3vSA (green), HST (blue); Area graphic: HSA (same value assigned to all residues in the "hot spot"); Normalized-Area graphic: normalized HSA for a 100-residue "hot spot" (NHSA). In the following rows we find the a^3vSA, the number of "hot spots" identified (nHS), the normalized number of "hot spots" for 100 residues (NnHS), the AAT, the THSA, the TA, the AAT and THSA divided by the number of residues (AATr and THSAr, respectively), and Na^4vSS. Finally, a row per residue is given with columns for the residue number, its one-letter code, a^4v, HSA, NHSA, and a^4vAHS (see additional file 3).

Results and Discussion

AGGRESCAN capabilities: Validation and Examples

Generation of protein aggregation profiles and prediction of aggregation "hot spots"

The prediction method implemented in AGGRESCAN has already allowed the identification of experimentally proved "hot spots" (HSs) in a set of both natively unfolded and globular pathogenic proteins: Aβ42 peptide, synuclein, amylin, prion protein, transthyretin, β2-microglobulin and lysozyme [12]. The main aims in the design of AGGRESCAN were the automation of this analysis for the study of large sets of polypeptide sequences, the introduction of new variables in the postprocessing of the aggregation profiles to provide a set of values that could be easily correlated with aggregation propensities and the presentation of results in a convenient and informative way. To further prove the general predictive ability of the method, the above-mentioned proteins, together with a new set of well studied protein sequences related to depositional diseases (aDan, aBri, apolipoproteins AI, AII, AIV, and CII, prolactin, insulin, Tau, fibrinogen, amyloid A, pulmonary surfactant protein, tropoelastin and medin), or shown to form amyloid *in vitro* (myoglobin, glycophorin A and amphoterin) have been analyzed with AGGRESCAN (Table 1). The predicted aggregation-prone protein regions have been validated by comparison to available experimental data on (i) regions known to promote aggregation, (ii) fragments known to aggregate *in vivo* (often after proteolysis) and (iii) synthetic short peptides shown to aggregate *in vitro* (references in Table 1). In the AGGRESCAN output, the sequence stretches with highest predicted aggregation propensity are shown in red in the peptide sequence column and appear as peaks in the Profile plots. The HS can be ranked according to their peak area (HSA) or normalized peak area (NHSA). Interestingly, protein segments that are experimentally known to be involved in aggregation are also found among the top ranked HS in their respective sequences based of the approach described here (Table 1), indicating that AGGRESCAN catches the main features underlying deposition in many conformational diseases. These results, together with previous experimental [10,17-20] and theoretical [21-24] data, suggest that specific short polypeptide stretches effectively promote and/or modulate protein amyloid formation.

One remarkable example in the test set is lung surfactant protein C (SP-C). This protein is expressed as a 197-amino acid proprotein that is processed to the 35-amino acid mature peptide. This fragment is associated with the development of pulmonary alveolar proteinosis (PAP). The bronchoalveolar fluid from PAP patients is rich in insoluble SP-C aggregates which exhibit the characteristic properties of amyloids by Congo red staining and electron microscopy. Moreover, the isolated peptide has been shown to form amyloid fibrils *in vitro* [25]. In good agreement with this data, AGGRESCAN predicts the SP-C region within the precursor as the HS with the highest aggregation propensity (Figure 1).

Other two interesting molecules are serum amyloid A (SAA) and Tau proteins, involved in systemic amyloidosis and Alzheimer's disease, respectively. AGGRESCAN detects only one HS in SAA and a very dominant one in Tau (Figure 1). In both cases, these sequence stretches correspond to the unique regions in SAA and Tau proved to be relevant for amyloidosis [26,27]. Importantly, the SAA and Tau sequences display highly negative Na^{4}vSS values, -28.2 and -32.5 respectively. Although this suggests an overall low aggregation propensity, the presence of specific HS that can act as nucleation points from which the ordered fibrillar structure can be expanded under certain circumstances, turn these proteins amyloidogenic. Actually, Tau is an usually highly soluble microtubule-associated protein [28] but in Alzheimer's disease it aggregates into fibres with a tendency to form neurofibrillary tangles. To date, only few 3D structures of amyloid assemblies at atomic resolution are available [29]. A crucial question is whether the formation of the tightly packed β-sheets observed in these structures is a generic backbone property or is dictated by the sequence. Interestingly enough, AGGRESCAN detects the presence of "hot spots" in most of the strands forming the intimate structure of the different protein fibrils (Table 2), providing additional support for the relevance of the primary structure on amyloid formation.

There are several computational approaches for detecting aggregation-prone regions and predicting polypeptide propensities for amyloid fibril formation. Some of them, including AGGRESCAN, rely on experimental or theoretical calculations of individual amino acid aggregation propensities and on the use of these values to scan protein sequences. The main difference between these algorithms is the way aggregation propensities are obtained. Pawar and co-workers proposed an aggregation scale based on phenomenological expressions relating protein intrinsic factors with the aggregation rates of a set point mutants scattered along acylphosphatase sequence and of a few other polypeptides [30]. As the fitting was done considering effects in both aggregation relevant and non-relevant regions, it is possible that the data do not necessarily reflect propensities within nucleating sequences. To address this point, Rojas Quijano and co-workers derived propensities from the analysis of the Tau-related amyloidogenic peptide Ac-VQIVYK-amide and its single site mutants Ac-VQIVXK-amide (X≠Cys) [19]. In AGGRESCAN, we somehow combine both approaches, in the sense that (i) propensities are calculated from the analysis of single mutants in a nucleating sequence (the central hydrophobic cluster of Aβ) which is perhaps the best well characterized aggregation-prone sequence in the literature and one of the few for which a high-resolution structure in the amyloid conformation is available, and (ii) we consider it in the context of the full length polypeptide (in fact fused to GFP, which acts as aggregation reporter) and not in an isolated manner as a short peptide. In addition, to the best of our knowledge our method is the only one in which aggregation propensities have been derived from experiments inside the cell, where the presence of the folding machinery might modulate the aggregation tendencies of polypeptides. Besides these three experimentally calculated propensity scales, Galzitskaya and co-workers have used the mean packing density for natural amino acid residues in protein structures, as a scale to predict amyloidogenic regions in proteins [31]. A comparative analysis of the four different scales shows that, despite these differences, there is a striking correlation between our in vivo obtained amino acid aggregation propensities and the others (Table 3), probably because they reflect a combination of properties characteristic of protein aggre-

Table 1: List and ranking of the predicted aggregation-prone regions in the different disease-linked polypeptides analyzed in this study and comparison with the available experimental data.

Protein	Experimental region[a]	Predicted region[b]	Ranking[c]	References
Abri	1–34	4–9	2/2	[52]
		15–28	1/2	
Adan	1–34	4–9	2/2	[53]
		15–24	1/2	
α-Synuclein	68–78	66–77	1/6	[54–56]
	31–109	36–42	2/6	[56]
		49–55	4/6	
		87–94	5/6	
Amphoterin	12–27	14–22	2/3	[57]
Amyloid-β-protein	17–21	17–22	2/2	[58]
	31–36/38–42	30–42	1/2	
Apoliprotein A-I	1–83	13–21	2/2	[59]
Apoliprotein A-II	N-terminal fragments	1–19	1/3	[11]
Apoliprotein A-IV	N-terminal fragments	1–19	1/6	[11]
Apoliprotein C-II	57–74	60–67	2/3	[60]
		69–76	1/3	
β2-Microgobulin	21–41	22–30	2/2	[61]
	59–79	59–70	1/2	[62]
Exon 30 Tropoelastin	1–25	1–7	2/2	[63]
		9/18	1/2	
Fibrinogen A α-chain	501–506	499–521	1/6	[64]
	482–504	501–506	1/5	
Glycophorin A	70–98	74–98	1/4	[65]
Insulin	1–38	12–19	1/3	[66]
		21–27	3/3	
Islet amyloid polypeptide	8–20	13–18	1/2	[67]
	20–29	24–28	2/2	[68]
Lysozyme (Hen)	40–64	54–62	2/4	[69]
	49–101	76–84	3/4	[70]
Medin	47–54	49–55	1/3	[71]
Myoglobin (Horse)	101–118	101–115	1/4	[72]
Prion Protein	106–147	117–136	3/6	[73]
		138–142	6/6	
Prolactin	1–34	10–32	2/9	[74]
Pulmonary surfactant protein	24–58	31–59	1/5	[25]
Serum Amyloid A	2–12	1–9	1/2	[75]
Tau	301–320	304–311	1/2	[27]
Transthyretin	10–20	12–19	2/7	[76]
	105–115	105–112	3/7	[77]
		114–123	4/7	

[a] Sequence stretches experimentally identified as critical for protein aggregation.
[b] Coincident aggregation-prone segments as predicted by AGGRESCAN.
[c] The rank position refers to the entire protein and reflects the importance of this specific "hot spot" (HS) relative to all the aggregation-prone regions identified by AGGRESCAN in the protein. (i.e., 1/4 indicates that this HS has the highest aggregation propensity of the four detected in a particular sequence by the software)

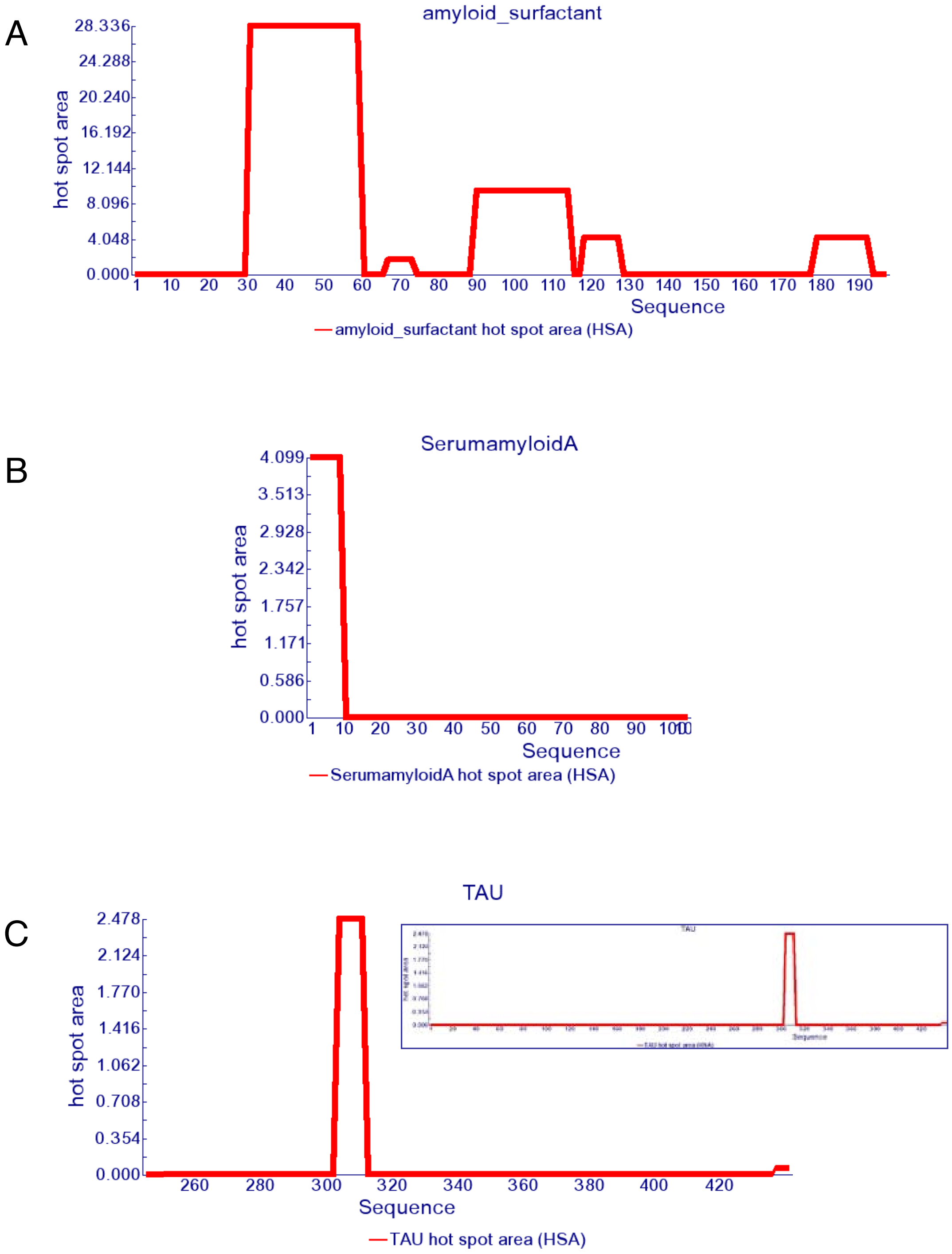


Figure 1
Hot spot area graphics. Hot spot area plots for a) lung surfactant protein C, b) serum amyloid A protein and c) Tau protein.

Table 2: Comparison of AGGRESCAN predictions with the structural composition of different amyloid fibrils.

Protein	Structure (β-strands)	Prediction	Reference
Aβ1-40	β 1: 12–24	17–22	[78]
	β 2: 30–40	30–40	
Amylin	β 1: 12–17	13–18	[79]
	β 2: 22–27	24–28	
	β 3: 31–37	-	[80]
HET's Prion	β 1: 226–234	-	
	β 2: 237–245	238–248	
	β 3: 262–270	263–267	
	β 4: 273–282	272–276	
Mouse Prion (89–143)	β 1: 112–124	115–129	[81]
β2- microglobulin (20–41)	β 1: 21–30	22–30	[82]
	β 2: 33–40	-	
Transthyretin (105–115)	β 1: 105–115	105–112	[83]

gation, such as hydrophobicity, secondary structure propensity or packing density. Importantly, although our method was not aimed at the specific identification of short amyloidogenic peptides, but rather of aggregation-prone sequences within natural proteins, AGGRESCAN identifies the presence of at least one hot spot in more than 80% of the amyloid forming sequences in a set of experimentally characterized peptide fragments of amyloidogenic proteins [32]. Also, using a database of six-residue peptides containing both amyloid formers and non-formers [32,33] the receiver operator characteristic (ROC) curve for our method compares well with those obtained using structure-based data, such us packing density on protein structures or the 3D profile method, based on the threading of six-residue peptides through the known crystal structure of the cross-β spine formed by the peptide NNQQNY from Sup35 yeast prion [32] (Figure 2).

Overall, the success of different computational approaches in predicting aggregation-prone regions allows to propose that aggregation propensity in polypeptide chains is ultimately dictated by the sequence. As it happens with the native conformation of proteins, the sequence contains intrinsic information that is relevant for the regular structural arrangement within β-aggregates, implying that the mechanism of amyloid fibril formation is similar for different peptides and proteins.

Prediction of the effects of protein mutation on the aggregation propensity

Aggregation propensity varies sensibly with the composition and especially the sequence of the polypeptide, in such a way that single amino acid substitutions in proteins associated to depositional diseases result in many cases in changes in the global protein aggregation propensity and sometimes lead to premature or acute pathological symptoms. Predicting the effect of a mutation on the aggregation tendency of a protein could help to anticipate the implications of that mutation in disease development or assist the design, production and storage of more soluble variants of biotechnologically relevant proteins and peptides [34].

Several AGGRESCAN output variables can be used to predict the effect of sequence variations on the aggregation propensities of a given polypeptide. The change in the normalized a^4v sum (Na4vSS) and Total Area (TA) are obvious indicators of changes in aggregation properties of the complete sequence due to point mutations. Nevertheless, a mutation will not always affect significantly the glo-

Table 3: Correlation coefficients (R) between the individual amino acid aggregation propensities used by AGGRESCAN and those used by other predictive methods.

	AGGRESCAN	AMYLOID1[a]	AMILOYD2[b]	AMYLOID3[c]
AGGRESCAN	*	0.849	0.794	0.867
AMYLOID1[a]	0.849	*	0.764	0.837
AMILOYD2[b]	0.794	0.764	*	0.807
AMYLOID3[c]	0.867	0.837	0.807	*

[a] AMYLOID1 corresponds to the method described in Ref. [22]
[b] AMYLOID2 corresponds to the method described in Ref. [19]
[c] AMYLOID3 corresponds to the method described in Ref. [31]

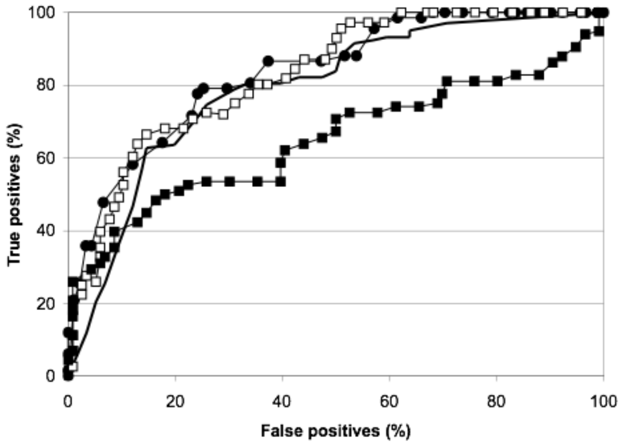


Figure 2

Comparative prediction performance of AGGRESCAN and structure-based methods. Comparative predictions of AGGRESCAN (solid circles), packing density profile [31] (no symbols), 3D Profile [32] using the NNQQNY template (solid squares) and 3D Profile using an ensemble of templates (empty squares). Predictions were tested in a Database of Fibril Formers and Non-Formers hexa-peptides. Predictions are shown as receiver-operator characteristic curves.

bal profile and changes in the number of HS (nHS), in the area over the HS threshold (AAT) or in the area assigned to the HS regions (THSA), are also informative. The normalized values (relative to the number of residues) AATr, THSAr and NHSA should be used if mutations resulting in sequence deletions or insertions are considered. To asses the capability of AGGRESCAN to predict sequence-variation effects we compared the experimentally observed aggregation changes reported in the literature for a group of more than 50 protein mutations with the change in different AGGRESCAN output variables. The analysis indicates that Na^{4}vSS is a good predictor for the effect on aggregation propensity changes in the polypeptide sequence on aggregation propensity (Table 4). The user has to take into account that a given mutation in a short protein is expected to have higher impact on aggregation that the same change in a longer sequence, where its effect can be more easily modulate by the rest of the sequence. These considerations are already included in the calculation of the Na4vSS values.

The algorithm predicts accurately a large set of natural and designed mutations of Aβ42 (the central hydrophobic region of this peptide was used for the derivation of the current a^{3}v parameter set of AGGRESCAN) (Table 4). As an example, Figure 3 shows how the F19T mutation, which strongly decreases the deposition of Aβ42 [35], results in the loss of the central HS in this peptide. Interestingly, it also anticipates the lower aggregation propensity of Aβ40 and the recent observation that longer Aβisoforms possess increased aggregation propensities [36]. Several natural occurring mutations have also been shown to affect the aggregation rate of Tau [37-40]. The predicted changes in the respective Na^{4}vSS correlate well with the experimental changes observed in these Tau variants (Table 4). Figure 3 shows the Area plot of wild type Tau and two of its mutants with highest, experimentally tested, aggregation propensities. The P301L substitution increases by 1,4 fold the area associated to the main "hot spot" in Tau. In addition, AGGRESCAN predicts the presence of a new HS in the S320F mutant, absent in the wild type form. This mutant is linked to tauthopaty, in which Tau accumulates in inclusion bodies [40].

Other disease-related protein mutants studied here are the recently described G4R and R68Stop of human Stefin B protein. These mutants have been related with the development of Myoclonus epilepsy of type 1. It has been described that R68Stop is more prone to aggregate than wild type Stefin, while the G4R mutant shows an opposite behavior, with a slower fibril formation rate [41]. In agreement with these experimental observations the algorithm predicts an increase in the Na^{4}vSS associated to the R68Stop mutation and a decrease for the Gly4Arg change (Table 4).

Type 1 serum amyloid A protein (SAA1) is associated with Familial Mediterranean fever (FMF). FMF patients' genotypes are thought to correlate with the different phenotypes of the disease. A recent study [42] concludes that the gamma SAA1 allele is more frequently observed in the population devoid of amyloidosis, thus suggesting a protective effect of this allele on the development of the illness. In agreement with these results the AGGRESCAN analysis of amyloid A sequence variants predicts that the gamma variant misses a HS and has a lower Na^{4}vSS than other alleles.

The Src homology 3 (SH3) domain of the p58 subunit of phosphatidyl-inositol-3 -kinase (PI3-SH3) is one of the best-characterized examples of a small globular protein unrelated to any known pathological condition that can form amyloid fibrils *in vitro* [43]. Aggregated species obtained from this protein have been found to be cytotoxic when added to cell cultures [44]. We have previously shown that the α-spectrin-SH3 (SPC-SH3) domain, which shares the same fold and 24% sequence identity with PI3-SH3, is a soluble protein that does not form amyloid fibrils under any conditions tested [45]. Nevertheless, a recent work found that the N47A mutation at the distal loop induces the formation of amyloid fibrils [46]. In contrast, the mutation of residue 47 to Gly does not promote

Table 4: Comparison of the predicted and experimentally tested effects of mutations on the aggregation propensity of amyloidogenic proteins.

Sequence Name	ΔNa^{4}vSS[a]	Experimental[b]	References
Peptide Aβ42 A21G	-16	-	[84]
Peptide Aβ42 E22K	15	+	[84]
Peptide Aβ42 E22G	29	+	[84]
Peptide Aβ42 E22Q	5	+	[84]
Peptide Aβ42 F19P	-68	-	[85]
Peptide Aβ42 F19T	-63	-	[35]
Peptide Aβ42 D23N	16	+	[86]
Peptide Aβ42 F19D	-118	-	[12]
Peptide Aβ42 I31L	-15	-	[87]
Peptide Aβ42 I32L	-15	-	[87]
Peptide Aβ42 I41G	-62	-	[87]
Peptide Aβ42 I41A	-49	-	[87]
Peptide Aβ42 I41L	-12	-	[87]
Peptide Aβ42 A42G	-10	-	[87]
Peptide Aβ42 A42V	32	+	[87]
Peptide Aβ42 Δ 1–4	59	+	[88]
Peptide Aβ42 Δ 1–9	237	+	[88]
Peptide Aβ42 Δ 40–42	-63	-	[88]
Peptide AβgΔ 41–42	-34	-	[36]
Peptide Aβg5	89	+	[36]
Peptide Aβg6	111	+	[36]
Peptide Aβg7	167	+	[36]
Peptide Aβ42 V12E+V18E+M35T+I41N	-312	-	[87]
Peptide Aβ42 F19S+L34P	-123	-	[87]
TAU R5L	2	+	[89]
TAU R406W	2	+	[90]
TAU G272V	2	+	[90]
TAU Y310W	0	=	[39]
TAU P301L	1	+	[40]
TAU S320F	2	+	[91]
α-synucleinA30P	-1	=	[92]
α-synucleinE46K	2	+	[92]
α-synucleinA53T	-1	+	[92]
α-synucleinA76E	-5	-	[93]
α-synucleinA76R	-3	-	[93]
Amylin (Rat) R18H	9	+	[94]
Amylin (Rat) L23F	17	+	[94]
Amylin (Rat) V26I	11	+	[94]
Amylin (Rat) R18H+L23F+V26I	34	+	[94]
Amylin (human) (22–27) N22A	21	+	[68]
Amylin (human) (22–27) F23A	-59	-	[68]
Amylin (human) (22–27)G24A	16	+	[68]
Amylin (human) (22–27) I26A	-61	-	[68]
Amylin (human) (22–27) L27A	-23	-	[68]
Amylin (human) S20G	-106	+	[95]
Amylin (human) ProIAPP	-90	+?	[96]
Human PrP H111A	5	+/=	[97]
Human PrP H111K	0	-/=	[97]
Human PrP A117V	7	+	[97]
Human PrP V210I	1	+	[98]
Stefin R68X	37	+	[41]
Stefin G4R	-6	-	[41]
SH3 n47a	17	+	[46]

[a]Relative change in Na^{4}vSS upon mutation, expressed as percentage.
ΔNa^{4}vSS = ((Na^{4}vSS$_{mut}$ - Na^{4}vSS$_{wt}$)/|Na^{4}vSS$_{wt}$|)*100
Na^{4}vSS$_{mut}$ refers to the Na^{4}vSS value of the mutant sequence.
Na^{4}vSS$_{wt}$ refers to the Na^{4}vSS value of the wild type sequence.
[b]Changes in aggregation determined experimentally.
Symbols: + increase; - decrease; = no significant change.

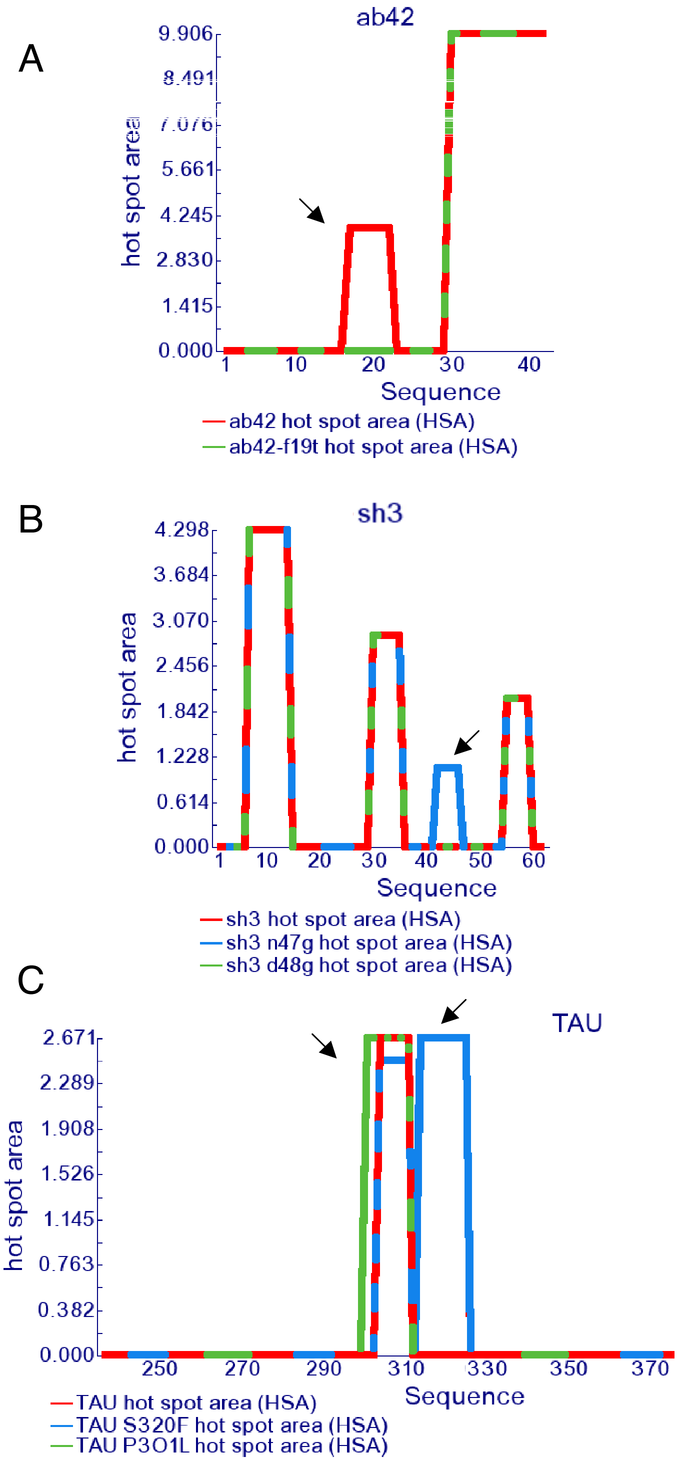


Figure 3

Changes in the hot spot area plot caused by point mutations in amyloidogenic proteins. a) Aβ42 wild type (red) and Aβ42 F19T mutant (green). b) SH3 wild type (red), SH3 D48G (green) and SH3 N47G (blue). c) TAU wild type (red), TAU P301L (green) and TAU S320F (blue).

aggregation (Ventura, S., unpublished results). According to AGGRESCAN a new HS occurs in the amyloidogenic mutant relative to both the wild type and N47G species, which could be responsible for its increased aggregation abilities (Figure 3).

Analysis of protein datasets

Besides analyzing the theoretical aggregation properties of single molecules and their individual mutants, AGGRESCAN is also able to deal simultaneously with a large number of sequences. This ability can be specially useful to compare the global aggregation properties of different protein sets and may help to delineate general rules underlying the relationship between the primary structure of proteins and peptides and their specific *in vivo* and *in vitro* depositional properties. With this aim we studied the correlation between the structural/aggregational features of 5 different groups of proteins and the predictions provided by AGGRESCAN. These datasets were: 1) natively globular proteins (160 proteins) (from SCOP, the ASTRAL40 set); 2) natively intrinsically unstructured proteins (51 proteins); 3) proteins which are soluble when overexpressed in bacteria (38 proteins); 4) proteins forming inclusion bodies when overexpressed in bacteria (121 proteins) and 5) amyloidogenic proteins (57 proteins) (see additional file 4).

When average AGGRESCAN output values are calculated and subsequently compared between data sets, it appears that the different protein groups can be individualized (Table 5), providing insight into the sequential determinants of protein aggregation and solubility. In this way, intrinsically unstructured proteins (IUP) clearly present the lowest output values of all datasets, in correlation with the general observation that unstructured proteins are usually resistant to aggregation and remain soluble after heat-treatment of the cells. Natively unfolded proteins exhibit a Na^{4}vSS value 7 times lower than that corresponding to the set of globular proteins from SCOP. Also, the normalized number of HS (NnHS) or the area over the threshold (AAT) and total HS area (THSA) are around 2 times higher in globular proteins than in IUP, showing that, in agreement with other automated analyses [47], the number of aggregation-prone sequence stretches is lower in IUP than in structured proteins. This result may reflect a negative natural selection against aggregation promoting residues and regions in IUP, where any HS will be exposed to solvent and accessible for the establishment of inter-molecular contacts that may finally lead to the build-up of aggregates. For the same reason, nature is likely to have provided globular proteins with a stable native conformation in which aggregation-prone sequences are buried in the inner hydrophobic core or involved in intra-molecular interactions [13,18]. This appears to be a successful evolutive strategy to avoid deposition, since few proteins aggregate from their folded state. Hence, amyloidogenic mutations in globular proteins usually result in destabilization of the native state, permitting exposure of natively hidden HS.

It has been recently shown that proteins that form inclusion bodies (IB) upon recombinant overexpression in *E. coli* and proteins that form amyloids *in vivo* and/or *in vitro* share a good number of structural and functional features, including high purity of the aggregates, enrichment in beta-sheet structure, amyloid-tropic dye binding or enhanced proteolytic resistance [3]. Comparison of the two protein sets in search for similar trends in the predictions showed that, unexpectedly, the AGGRESCAN values for amyloid forming proteins are closer to those for IUP than for any other of the analysed datasets. Amyloid proteins have a lower Na^{4}vSS and less HS than proteins in the

Table 5: Comparison of the different AGGRESCAN parameters for globular, natively unstructured, amyloidogenic, soluble and insoluble proteins.

Set Name	Globular[1]	Unfolded[2]	Amyloid[3]	IBs[4]	Soluble[5]
a3vSA	***-0.04***	***-0.28***	***-0.12***	***-0.02***	***-0.05***
nHS	9.54	5.63	5.86	11.97	10.34
NnHS	***3.89***	***2.06***	***2.89***	***3.50***	***3.35***
AAT	29.94	18.21	24.51	41.27	34.43
THSA	25.58	14.97	21.26	36.00	29.61
TA	-5.17	-60.95	-26.42	-5.00	-5.55
AATr	***0.12***	***0.07***	***0.13***	***0.13***	***0.12***
THSAr	***0.11***	***0.05***	***0.11***	***0.11***	***0.09***
Na4vSS	***-4.26***	***-28.73***	***-12.96***	***-2.51***	***-5.18***

In bold and italics are shown those parameters that are normalized by the number of residues, allowing direct comparison of datasets independently of protein size.

[1] Natively globular proteins: 160 proteins randomly selected from SCOP (the ASTRAL40 set)

[2] Natively intrinsically unstructured proteins: 51 proteins

[3] Amyloidogenic proteins: 57 proteins

[4] Proteins forming inclusion bodies when overexpressed in bacteria: 121 proteins

[5] Proteins which are soluble when overexpressed in bacteria: 38 proteins

IB or globular SCOP dataset (Figure 4). In contrast, the HSs in amyloid proteins comprise an area similar to those in IB or globular proteins, which is, however, significantly higher than the average HS area in IUP. These results suggests that, globally, the sequences of amyloidogenic proteins, like those of IUPs, have a low aggregation propensity, although the existence of specific aggregation-prone regions, absent or minor in IUPs, in a context in which they can act as specific and obligatory nucleation points from which the fibrillar structure could be expanded, finally results in highly ordered aggregates (Figure 4). This would explain why point mutations in the HSs of amyloidogenic proteins have usually a huge impact in their solubility, as they would modify the properties of one of the few points in the sequence that can promote and/or modulate aggregation. In contrast, the paradoxically higher-ranking global aggregation propensity of IB protein sequences is likely to indicate that here HS would play a less important role, since aggregation can also occur non specifically from many regions in the protein sequence. This would result in less structured deposits, and would also explain the rather moderate role of point mutations in IB aggregate formation. In other words, a given HS would promote specific amyloid formation in a low aggregating background, as its aggregation tendency outstands from the rest of the sequence. Conversely, the same HS needs to compete with the rest of the sequence to nucleate aggregation in a highly aggregating context (Figure 5). For the same reason unstructured aggregation is usually a much faster event than amyloidogenesis. Recent theoretical and experimental data support this view by showing that prevention of aggregation does not necessarily mean that amyloid fibril formation is abolished and *vice versa* [48]. This indicates that, despite the fact that aggregates and amyloid fibrils share many features, and the protein regions involved in their formation presumably intersect, they probably differ in the number and specificity of intermolecular contacts involved in the nucleation and stabilization of both types of polypeptide associations.

Recombinant protein production is an essential tool for the biotechnological industry and supports expanding areas of basic and biomedical research, including structural genomics and proteomics. The solubility of proteins expressed in bacteria under mass-production conditions is of major concern, since many recombinant polypeptides produced in bacteria accumulate as insoluble, often refractile, aggregates known as inclusion bodies (IBs) [49], excluding many biotechnologically relevant protein species from the market due to economically inconvenient yields. To date, the solubility of a given gene product has not been anticipated before gene expression. The comparison between the AGGRESCAN output values for proteins shown to be soluble under overexpression conditions in *E. coli* and those forming inclusion bodies shows that they share a similar number of HSs per 100 residues (NnHS), an expected output when considering that most proteins in both datasets are globular. However, IB-forming proteins have higher Na^{4}vSS values than soluble proteins, suggesting that soluble proteins have, on average, a lower intrinsic aggregation tendency than IB-forming species, which may determine, at least partially, their relative behaviour upon overexpression. Overall, the predicted aggregation of proteins in the SCOP database is intermediate between that of soluble and insoluble proteins, suggesting that, in agreement with experimental observations, only a part of them would remain in the soluble cell fraction upon recombinant production. Although AGGRESCAN is able to catch the average trends in the aggregation of IBs and soluble protein groups, the individual outputs for proteins from both groups overlap significantly, making the prediction of the recombinant behaviour of a given sequence difficult in its present form. Besides, aggregation during recombinant production is the net result of several extrinsic and intrinsic factors, their relative importance depending on the protein and expression contexts.

Conclusion

The software and web interface developed in the present study allow an easy and accurate identification and ranking of aggregation-prone regions in polypeptides. AGGRESCAN is also able to anticipate the effect of genetic or artificially introduced sequence changes on the aggregation properties of polypeptides. In addition to the investigation of the role of the primary sequence on protein aggregation and protein solubility, the algorithm can be used in the design of strategies for the treatment of amyloidogenesis, by targeting therapies to those regions in the polypeptide chain whose aggregation propensities outstand from the rest, provided that they are or become exposed to solvent in the disease-related protein conformation. The surprising observation that the aggregation propensities of amyloid sequences tend to be low, suggests that blocking the "hot spots" of aggregation in these proteins, either chemically or by mutation, may have a huge impact on their solubility. Interestingly enough, protein-protein interactions are often mediated through an energetic hot spot [50] which comprises few interface residues that contribute to most of the binding energy; identification and blocking of those sequence stretches has been suggested as an strategy to modulate protein interactions [51]. The ability of AGGRESCAN to analyze simultaneously the aggregation properties of large sets of protein sequences might be important for protein production in large-scale structural initiatives, for the analysis of the distribution of aggregation-prone regions in complete genomes or for evolutive studies, since it is likely that nat-

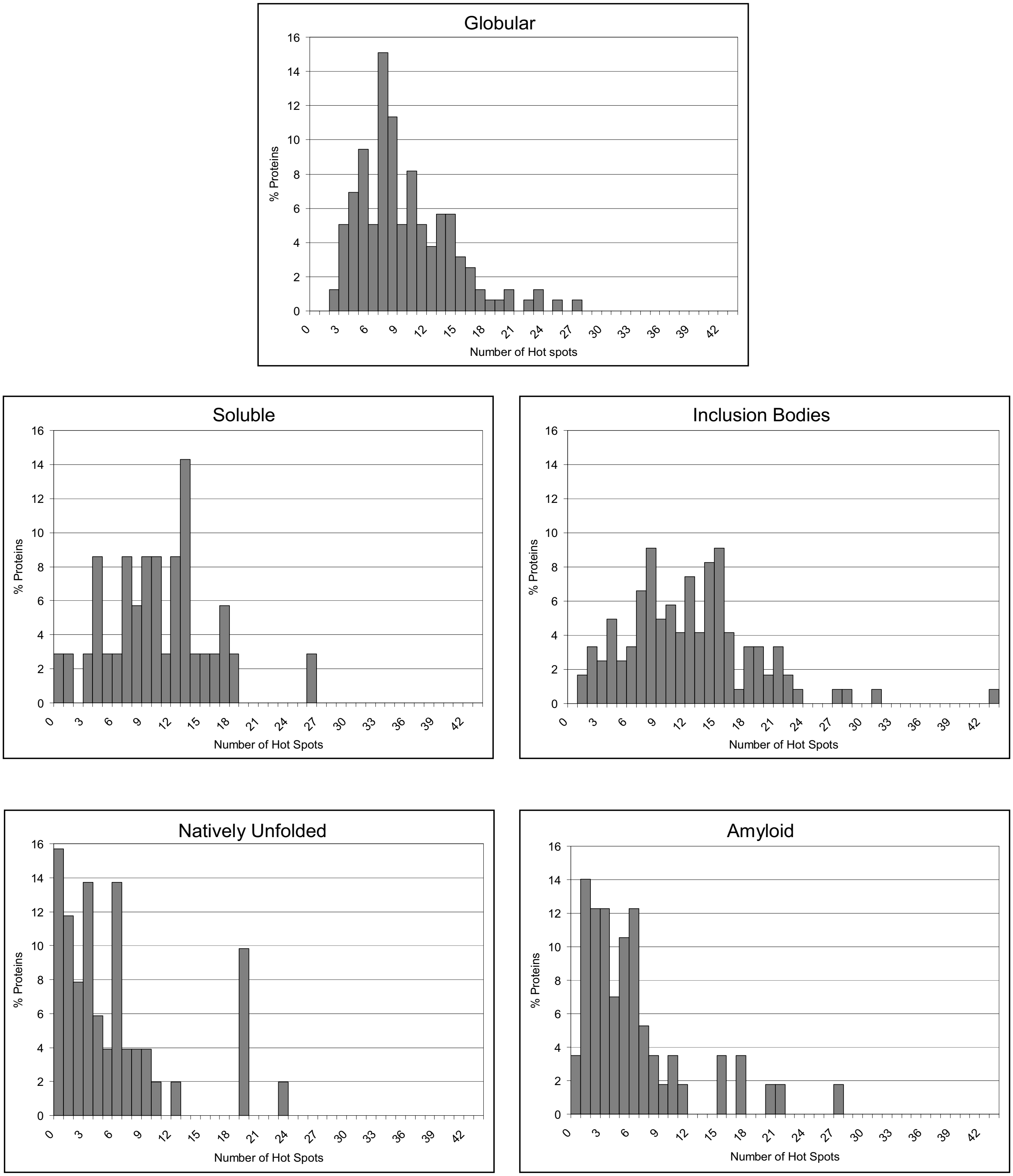


Figure 4

"Hot spots" distribution in different protein groups. Distribution of the number of "hot spots" relative to sequence length in the following protein datasets: natively globular proteins, intrinsically unstructured proteins, amyloidogenic proteins, soluble proteins when overexpressed in bacteria and proteins forming inclusion bodies when overexpressed in bacteria.

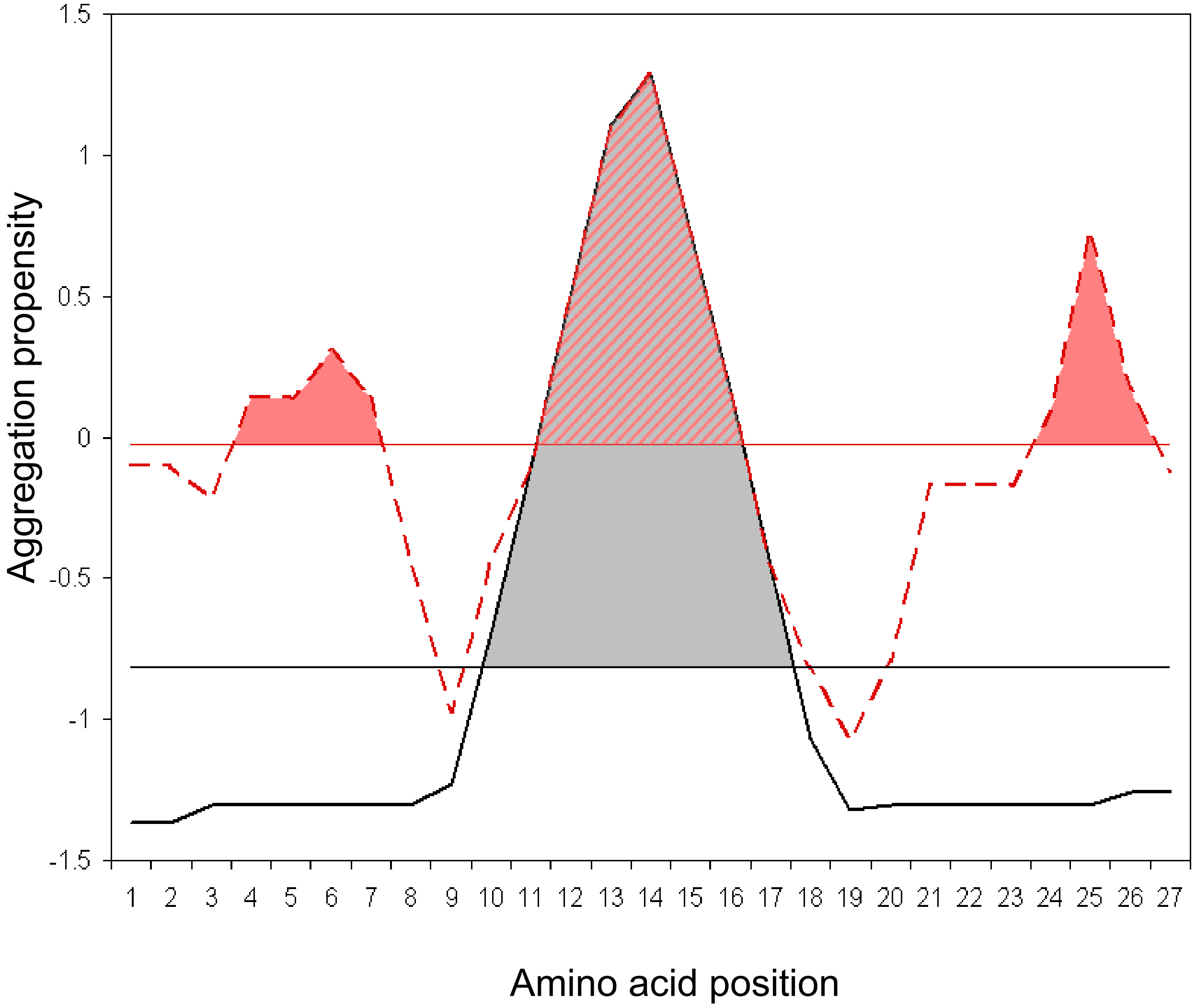


Figure 5

Modulation of hot spot nucleation specificity by global aggregation propensity. The black solid line represents a standard amyloidogenic protein aggregation profile, with only one "hot spot" and low global aggregation propensity. The pink discontinuous line corresponds to a typical aggregation profile from an inclusion-body-forming protein, with many "hot spots" and high global aggregation propensity. The horizontal lines represent the aggregation-propensity average thresholds for each sequence. The coloured regions indicate the area of each "hot spot" over the aggregation propensity threshold. It is proposed that a higher area over the threshold promotes a more specific aggregation reaction, resulting in highly ordered deposits.

ural protein sequences have evolved in part to code for avoidance of aggregation.

Availability and requirements

Project name: AGGRESCAN

Project home page: http://bioinf.uab.es/aggrescan/

Operating system(s): Platform independent

Programming language: a computing core coded in C and a front end written in a combination of html and perl cgi.

Other requirements: a web browser, such as Internet Explorer, Safari, or Firefox.

Any restrictions to use by non-academics: Incorporation into commercial products restricted.

Authors' contributions

OCS implemented the software, NSG analyzed and prepared the final data and figures. FXA and JV contributed to data interpretation and manuscript redaction. XD directed the implementation of the software and contributed to manuscript redaction. SV directed the work and prepared the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

AGGRESCAN aggregation propensities
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-65-S1.pdf]

Additional file 2

Help file of AGGRESCAN
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-65-S2.pdf]

Additional file 3

Example of an output of AGGRESCAN
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-65-S3.pdf]

Additional file 4

Protein data sets tested with AGGRESCAN
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-65-S4.pdf]

Acknowledgements

This work was supported by grants BIO2004-05879 and BIO2003-02848 (Ministerio de Educación y Ciencia, Spain) and SGR2005-00037 and SGR2005-01037 (Generalitat de Catalunya, Spain). NS is recipient of a fellowship from the Spanish Ministerio de Educación y Ciencia.

References

1. Fink AL: **Protein aggregation: folding aggregates, inclusion bodies and amyloid.** *Fold Des* 1998, **3:**R9 -23.
2. Smith A: **protein misfolding.** *Nature* 2003, **426:**883 -8883.
3. Ventura S, Villaverde A: **Protein quality in bacterial inclusion bodies.** *Trends Biotechnol* 2006, **24(4):**179-185.
4. Treuheit MJ, Kosky AA, Brems DN: **Inverse relationship of protein concentration and aggregation.** *Pharm Res* 2002, **19(4):**511-516.
5. Dobson CM: **Protein-misfolding diseases: Getting out of shape.** *Nature* 2002, **418:**729 -7730.
6. Cohen FE, Kelly JW: **Therapeutic approaches to protein-misfolding diseases.** *Nature* 2003, **426:**905 -9909.
7. Rochet JC, Lansbury PT: **Amyloid fibrillogenesis: themes and variations.** *Curr Opin Struct Biol* 2000, **10:**60 -668.
8. Stefani M, Dobson CM: **Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution.** *J Mol Med* 2003, **81(11):**678-699.
9. Ivanova MI, Sawaya MR, Gingery M, Attinger A, Eisenberg D: **An amyloid-forming segment of {beta}2-microglobulin suggests a molecular model for the fibril.** *PNAS* 2004, **101(29):**10584-10589.
10. Ventura S, Zurdo J, Narayanan S, Parreno M, Mangues R, Reif B, Chiti F, Giannoni E, Dobson CM, Aviles FX, Serrano L: **Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case.** *Proc Natl Acad Sci U S A* 2004, **101:**7258 -77263.
11. Chiti F, Dobson CM: **Protein misfolding, functional amyloid, and human disease.** *Annu Rev Biochem* 2006, **75:**333-366.
12. de Groot NS, Aviles FX, Vendrell J, Ventura S: **Mutagenesis of the central hydrophobic cluster in Abeta42 Alzheimer's peptide. Side-chain properties correlate with aggregation propensities.** *Febs J* 2006, **273(3):**658-668.
13. de Groot N, Pallares I, Aviles F, Vendrell J, Ventura S: **Prediction of "hot spots" of aggregation in disease-linked polypeptides.** *BMC Structural Biology* 2005, **5(1):**18.
14. Chiti F, Stefani M, Taddei N, Ramponi G, Dobson CM: **Rationalization of the effects of mutations on peptide and protein aggregation rates.** *Nature* 2003, **424(6950):**805-808.
15. [http://www.expasy.org/tools/pscale/A.A.Swiss-Prot.html].
16. Williams AD, Portelius E, Kheterpal I, Guo JT, Cook KD, Xu Y, Wetzel R: **Mapping abeta amyloid fibril secondary structure using scanning proline mutagenesis.** *J Mol Biol* 2004, **335(3):**833-842.
17. Chiti F, Webster P, Taddei N, Clark A, Stefani M, Ramponi G, Dobson CM: **Designing conditions for in vitro formation of amyloid protofilaments and fibrils.** *Proc Natl Acad Sci U S A* 1999, **96(7):**3590-3594.
18. Chiti F, Calamai M, Taddei N, Stefani M, Ramponi G, Dobson CM: **Studies of the aggregation of mutant proteins in vitro provide insights into the genetics of amyloid diseases.** *Proc Natl Acad Sci U S A* 2002, **99 Suppl 4:**16419-16426.
19. Rojas Quijano FA, Morrow D, Wise BM, Brancia FL, Goux WJ: **Prediction of nucleating sequences from amyloidogenic propensities of tau-related peptides.** *Biochemistry* 2006, **45(14):**4638-4652.
20. Ivanova MI, Thompson MJ, Eisenberg D: **A systematic screen of beta(2)-microglobulin and insulin for amyloid-like segments.** *Proc Natl Acad Sci U S A* 2006, **103(11):**4079-4082.
21. Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L: **Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins.** *Nat Biotechnol* 2004, **22:**1302 -11306.
22. DuBay KF, Pawar AP, Chiti F, Zurdo J, Dobson CM, Vendruscolo M: **Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains.** *J Mol Biol* 2004, **341(5):**1317-1326.
23. Tartaglia GG, Cavalli A, Pellarin R, Caflisch A: **Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences.** *Protein Sci* 2005, **14(10):**2723-2734.

24. Idicula-Thomas S, Balaji PV: **Understanding the relationship between the primary structure of proteins and their amyloidogenic propensity: clues from inclusion body formation.** *Protein Eng Des Sel* 2005, **18(4):**175-180.
25. Johansson J, Weaver TE, Tjernberg LO: **Proteolytic generation and aggregation of peptides from transmembrane regions: lung surfactant protein C and amyloid beta-peptide.** *Cell Mol Life Sci* 2004, **61(3):**326-335.
26. Westermark P, Johnson KH, O'Brien TD, Betsholtz C: **Islet amyloid polypeptide--a novel controversy in diabetes research.** *Diabetologia* 1992, **35(4):**297-303.
27. Margittai M, Langen R: **Template-assisted filament growth by parallel stacking of tau.** *Proc Natl Acad Sci U S A* 2004, **101(28):**10278-10283.
28. Selkoe DJ: **Cell biology of protein misfolding: the examples of Alzheimer's and Parkinson's diseases.** *Nat Cell Biol* 2004, **6(11):**1054-1061.
29. Nelson R, Eisenberg D: **Structural models of amyloid-like fibrils.** *Adv Protein Chem* 2006, **73:**235-282.
30. Pawar AP, Dubay KF, Zurdo J, Chiti F, Vendruscolo M, Dobson CM: **Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases.** *J Mol Biol* 2005, **350(2):**379-392.
31. Galzitskaya OV, Garbuzynskiy SO, Lobanov MY: **Prediction of amyloidogenic and disordered regions in protein chains.** *PLoS Comput Biol* 2006, **2(12):**e177.
32. Thompson MJ, Sievers SA, Karanicolas J, Ivanova MI, Baker D, Eisenberg D: **The 3D profile method for identifying fibril-forming segments of proteins.** *Proc Natl Acad Sci U S A* 2006, **103(11):**4074-4078.
33. Lopez De La Paz M, Goldie K, Zurdo J, Lacroix E, Dobson CM, Hoenger A, Serrano L: **De novo designed peptide-based amyloid fibrils.** *Proc Natl Acad Sci U S A* 2002, **99(25):**16052-16057.
34. Fowler SB, Poon S, Muff R, Chiti F, Dobson CM, Zurdo J: **Rational design of aggregation-resistant bioactive peptides: reengineering human calcitonin.** *Proc Natl Acad Sci U S A* 2005, **102(29):**10105-10110.
35. Esler WP, Stimson ER, Ghilardi JR, Lu YA, Felix AM, Vinters HV, Mantyh PW, Lee JP, Maggio JE: **Point substitution in the central hydrophobic cluster of a human beta-amyloid congener disrupts peptide folding and abolishes plaque competence.** *Biochemistry* 1996, **35:**13914 -13921.
36. Lambermon MH, Rappaport RV, McLaurin J: **Biophysical characterization of longer forms of amyloid beta peptides: possible contribution to flocculent plaque formation.** *J Neurochem* 2005, **95(6):**1667-1676.
37. Gamblin TC, Berry RW, Binder LI: **Tau polymerization: role of the amino terminus.** *Biochemistry* 2003, **42(7):**2252-2257.
38. Barghorn S, Mandelkow E: **Toward a unified scheme for the aggregation of tau into Alzheimer paired helical filaments.** *Biochemistry* 2002, **41(50):**14885-14896.
39. Li L, von Bergen M, Mandelkow EM, Mandelkow E: **Structure, stability, and aggregation of paired helical filaments from tau protein and FTDP-17 mutants probed by tryptophan scanning mutagenesis.** *J Biol Chem* 2002, **277(44):**41390-41400.
40. Yao TM, Tomoo K, Ishida T, Hasegawa H, Sasaki M, Taniguchi T: **Aggregation analysis of the microtubule binding domain in tau protein by spectroscopic methods.** *J Biochem (Tokyo)* 2003, **134(1):**91-99.
41. Rabzelj S, Turk V, Zerovnik E: **In vitro study of stability and amyloid-fibril formation of two mutants of human stefin B (cystatin B) occurring in patients with EPM1.** *Protein Sci* 2005, **14(10):**2713-2722.
42. Delibas A, Oner A, Balci B, Demircin G, Bulbul M, Bek K, Erdogan O, Baysun S, Yilmaz E: **Genetic risk factors of amyloidogenesis in familial Mediterranean fever.** *Am J Nephrol* 2005, **25(5):**434-440.
43. Jimenez JL, Guijarro JI, Orlova E, Zurdo J, Dobson CM, Sunde M, Saibil HR: **Cryo-electron microscopy structure of an SH3 amyloid fibril and model of the molecular packing.** *Embo J* 1999, **18(4):**815-821.
44. Bucciantini M, Giannoni E, Chiti F, Baroni F, Formigli L, Zurdo J, Taddei N, Ramponi G, Dobson CM, Stefani M: **Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases.** *Nature* 2002, **416(6880):**507-511.
45. Ventura S, Lacroix E, Serrano L: **Insights into the origin of the tendency of the PI3-SH3 domain to form amyloid fibrils.** *J Mol Biol* 2002, **322:**1147 -11458.
46. Morel B, Casares S, Conejero-Lara F: **A single mutation induces amyloid aggregation in the alpha-spectrin SH3 domain: analysis of the early stages of fibril formation.** *J Mol Biol* 2006, **356(2):**453-468.
47. Linding R, Schymkowitz J, Rousseau F, Diella F, Serrano L: **A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins.** *J Mol Biol* 2004, **342(1):**345-353.
48. Rousseau F, Schymkowitz J, Serrano L: **Protein aggregation and amyloidosis: confusion of the kinds?** *Curr Opin Struct Biol* 2006, **16(1):**118-126.
49. Villaverde A, Carrio MM: **Protein aggregation in recombinant bacteria: biological role of inclusion bodies.** *Biotechnol Lett* 2003, **25(17):**1385-1395.
50. Clackson T, Wells JA: **A hot spot of binding energy in a hormone-receptor interface.** *Science* 1995, **267(5196):**383-386.
51. Keskin O, Ma B, Nussinov R: **Hot regions in protein--protein interactions: the organization and contribution of structurally conserved hot spot residues.** *J Mol Biol* 2005, **345(5):**1281-1294.
52. El-Agnaf O, Gibson G, Lee M, Wright A, Austen BM: **Properties of neurotoxic peptides related to the Bri gene.** *Protein Pept Lett* 2004, **11(3):**207-212.
53. El-Agnaf OM, Nagala S, Patel BP, Austen BM: **Non-fibrillar oligomeric species of the amyloid ABri peptide, implicated in familial British dementia, are more potent at inducing apoptotic cell death than protofibrils or mature fibrils.** *J Mol Biol* 2001, **310(1):**157-168.
54. Goedert M: **Alpha-synuclein and neurodegenerative diseases.** *Nat Rev Neurosci* 2001, **2(7):**492-501.
55. Bodles AM, Guthrie DJ, Greer B, Irvine GB: **Identification of the region of non-Abeta component (NAC) of Alzheimer's disease amyloid responsible for its aggregation and toxicity.** *J Neurochem* 2001, **78(2):**384-395.
56. Miake H, Mizusawa H, Iwatsubo T, Hasegawa M: **Biochemical characterization of the core structure of alpha-synuclein filaments.** *J Biol Chem* 2002, **277(21):**19213-19219.
57. Kallijarvi J, Haltia M, Baumann MH: **Amphoterin includes a sequence motif which is homologous to the Alzheimer's beta-amyloid peptide (Abeta), forms amyloid fibrils in vitro, and binds avidly to Abeta.** *Biochemistry* 2001, **40(34):**10032-10037.
58. Morimoto A, Irie K, Murakami K, Masuda Y, Ohigashi H, Nagao M, Fukuda H, Shimizu T, Shirasawa T: **Analysis of the secondary structure of beta-amyloid (Abeta42) fibrils by systematic proline replacement.** *J Biol Chem* 2004, **279(50):**52781-52788.
59. Nichols WC, Dwulet FE, Liepnieks J, Benson MD: **Variant apolipoprotein AI as a major constituent of a human hereditary amyloid.** *Biochem Biophys Res Commun* 1988, **156(2):**762-768.
60. Wilson LM, Mok YF, Binger KJ, Griffin MD, Mertens HD, Lin F, Wade JD, Gooley PR, Howlett GJ: **A Structural Core Within Apolipoprotein C-II Amyloid Fibrils Identified Using Hydrogen Exchange and Proteolysis.** *J Mol Biol* 2007, **366(5):**1639-51.
61. Hasegawa K, Ohhashi Y, Yamaguchi I, Takahashi N, Tsutsumi S, Goto Y, Gejyo F, Naiki H: **Amyloidogenic synthetic peptides of beta2-microglobulin--a role of the disulfide bond.** *Biochem Biophys Res Commun* 2003, **304(1):**101-106.
62. Jones S, Manning J, Kad NM, Radford SE: **Amyloid-forming peptides from beta2-microglobulin-Insights into the mechanism of fibril formation in vitro.** *J Mol Biol* 2003, **325(2):**249-257.
63. Tamburro AM, Pepe A, Bochicchio B, Quaglino D, Ronchetti IP: **Supramolecular amyloid-like assembly of the polypeptide sequence coded by exon 30 of human tropoelastin.** *J Biol Chem* 2005, **280(4):**2682-2690.
64. Hamidi Asl L, Liepnieks JJ, Uemichi T, Rebibou JM, Justrabo E, Droz D, Mousson C, Chalopin JM, Benson MD, Delpech M, Grateau G: **Renal amyloidosis with a frame shift mutation in fibrinogen aalpha-chain gene producing a novel amyloid protein.** *Blood* 1997, **90(12):**4799-4805.
65. Liu W, Crocker E, Zhang W, Elliott JI, Luy B, Li H, Aimoto S, Smith SO: **Structural role of glycine in amyloid fibrils formed from transmembrane alpha-helices.** *Biochemistry* 2005, **44(9):**3591-3597.

66. Jimenez JL, Nettleton EJ, Bouchard M, Robinson CV, Dobson CM, Saibil HR: **The protofilament structure of insulin amyloid fibrils.** *Proc Natl Acad Sci U S A* 2002, **99(14):**9196-9201.
67. Scrocchi LA, Ha K, Chen Y, Wu L, Wang F, Fraser PE: **Identification of minimal peptide sequences in the (8-20) domain of human islet amyloid polypeptide involved in fibrillogenesis.** *J Struct Biol* 2003, **141(3):**218-227.
68. Azriel R, Gazit E: **Analysis of the minimal amyloid-forming fragment of the islet amyloid polypeptide. An experimental support for the key role of the phenylalanine residue in amyloid formation.** *J Biol Chem* 2001, **276(36):**34156-34161.
69. Krebs MR, Wilkins DK, Chung EW, Pitkeathly MC, Chamberlain AK, Zurdo J, Robinson CV, Dobson CM: **Formation and seeding of amyloid fibrils from wild-type hen lysozyme and a peptide fragment from the beta-domain.** *J Mol Biol* 2000, **300(3):**541-549.
70. Frare E, Polverino De Laureto P, Zurdo J, Dobson CM, Fontana A: **A highly amyloidogenic region of hen lysozyme.** *J Mol Biol* 2004, **340(5):**1153-1165.
71. Reches M, Gazit E: **Amyloidogenic hexapeptide fragment of medin: homology to functional islet amyloid polypeptide fragments.** *Amyloid* 2004, **11(2):**81-89.
72. Fandrich M, Forge V, Buder K, Kittler M, Dobson CM, Diekmann S: **Myoglobin forms amyloid fibrils by association of unfolded polypeptide segments.** *Proc Natl Acad Sci U S A* 2003, **100(26):**15463-15468.
73. Tagliavini F, Prelli F, Verga L, Giaccone G, Sarma R, Gorevic P, Ghetti B, Passerini F, Ghibaudi E, Forloni G, *et al.*: **Synthetic peptides homologous to prion protein residues 106-147 form amyloid-like fibrils in vitro.** *Proc Natl Acad Sci U S A* 1993, **90(20):**9678-9682.
74. Hinton DR, Polk RK, Linse KD, Weiss MH, Kovacs K, Garner JA: **Characterization of spherical amyloid protein from a prolactin-producing pituitary adenoma.** *Acta Neuropathol (Berl)* 1997, **93(1):**43-49.
75. Westermark GT, Engstrom U, Westermark P: **The N-terminal segment of protein AA determines its fibrillogenic property.** *Biochem Biophys Res Commun* 1992, **182(1):**27-33.
76. Jarvis JA, Kirkpatrick A, Craik DJ: **1H NMR analysis of fibril-forming peptide fragments of transthyretin.** *Int J Pept Protein Res* 1994, **44(4):**388-398.
77. Jaroniec CP, MacPhee CE, Bajaj VS, McMahon MT, Dobson CM, Griffin RG: **High-resolution molecular structure of a peptide in an amyloid fibril determined by magic angle spinning NMR spectroscopy.** *Proc Natl Acad Sci U S A* 2004, **101(3):**711-716.
78. Petkova AT, Ishii Y, Balbach JJ, Antzutkin ON, Leapman RD, Delaglio F, Tycko R: **A structural model for Alzheimer's beta -amyloid fibrils based on experimental constraints from solid state NMR.** *Proc Natl Acad Sci U S A* 2002, **99(26):**16742-16747.
79. Kajava AV, Aebi U, Steven AC: **The parallel superpleated beta-structure as a model for amyloid fibrils of human amylin.** *J Mol Biol* 2005, **348(2):**247-252.
80. Ritter C, Maddelein ML, Siemer AB, Luhrs T, Ernst M, Meier BH, Saupe SJ, Riek R: **Correlation of structural elements and infectivity of the HET-s prion.** *Nature* 2005, **435(7043):**844-848.
81. Lim KH, Nguyen TN, Damo SM, Mazur T, Ball HL, Prusiner SB, Pines A, Wemmer DE: **Solid-state NMR structural studies of the fibril form of a mutant mouse prion peptide PrP89-143(P101L).** *Solid State Nucl Magn Reson* 2006, **29(1-3):**183-190.
82. Iwata K, Fujiwara T, Matsuki Y, Akutsu H, Takahashi S, Naiki H, Goto Y: **3D structure of amyloid protofilaments of beta2-microglobulin fragment probed by solid-state NMR.** *Proc Natl Acad Sci U S A* 2006, **103(48):**18119-18124.
83. Jaroniec CP, MacPhee CE, Astrof NS, Dobson CM, Griffin RG: **Molecular conformation of a peptide fragment of transthyretin in an amyloid fibril.** *Proc Natl Acad Sci U S A* 2002, **99(26):**16748-16753.
84. Yamamoto N, Hasegawa K, Matsuzaki K, Naiki H, Yanagisawa K: **Environment- and mutation-dependent aggregation behavior of Alzheimer amyloid beta-protein.** *J Neurochem* 2004, **90(1):**62-69.
85. Cannon MJ, Williams AD, Wetzel R, Myszka DG: **Kinetic analysis of beta-amyloid fibril elongation.** *Anal Biochem* 2004, **328(1):**67-75.
86. Van Nostrand WE, Melchor JP, Cho HS, Greenberg SM, Rebeck GW: **Pathogenic effects of D23N Iowa mutant amyloid beta -protein.** *J Biol Chem* 2001, **276(35):**32860-32866.
87. Wurth C, Guimard NK, Hecht MH: **Mutations that reduce aggregation of the Alzheimer's Abeta42 peptide: an unbiased search for the sequence determinants of Abeta amyloidogenesis.** *J Mol Biol* 2002, **319(5):**1279-1290.
88. Jarrett JT, Berger EP, Lansbury PT Jr.: **The carboxy terminus of the beta amyloid protein is critical for the seeding of amyloid formation: implications for the pathogenesis of Alzheimer's disease.** *Biochemistry* 1993, **32(18):**4693-4697.
89. Gamblin TC, Chen F, Zambrano A, Abraha A, Lagalwar S, Guillozet AL, Lu M, Fu Y, Garcia-Sierra F, LaPointe N, Miller R, Berry RW, Binder LI, Cryns VL: **Caspase cleavage of tau: linking amyloid and neurofibrillary tangles in Alzheimer's disease.** *Proc Natl Acad Sci U S A* 2003, **100(17):**10032-10037.
90. Barghorn S, Zheng-Fischhofer Q, Ackmann M, Biernat J, von Bergen M, Mandelkow EM, Mandelkow E: **Structure, microtubule interactions, and paired helical filament aggregation by tau mutants of frontotemporal dementias.** *Biochemistry* 2000, **39(38):**11714-11721.
91. Rosso SM, van Herpen E, Deelen W, Kamphorst W, Severijnen LA, Willemsen R, Ravid R, Niermeijer MF, Dooijes D, Smith MJ, Goedert M, Heutink P, van Swieten JC: **A novel tau mutation, S320F, causes a tauopathy with inclusions similar to those in Pick's disease.** *Ann Neurol* 2002, **51(3):**373-376.
92. Choi W, Zibaee S, Jakes R, Serpell LC, Davletov B, Crowther RA, Goedert M: **Mutation E46K increases phospholipid binding and assembly into filaments of human alpha-synuclein.** *FEBS Lett* 2004, **576(3):**363-368.
93. Giasson BI, Murray IV, Trojanowski JQ, Lee VM: **A hydrophobic stretch of 12 amino acid residues in the middle of alpha-synuclein is essential for filament assembly.** *J Biol Chem* 2001, **276(4):**2380-2386.
94. Green J, Goldsbury C, Mini T, Sunderji S, Frey P, Kistler J, Cooper G, Aebi U: **Full-length rat amylin forms fibrils following substitution of single residues from human amylin.** *J Mol Biol* 2003, **326(4):**1147-1156.
95. Sakagashira S, Sanke T, Hanabusa T, Shimomura H, Ohagi S, Kumagaye KY, Nakajima K, Nanjo K: **Missense mutation of amylin gene (S20G) in Japanese NIDDM patients.** *Diabetes* 1996, **45(9):**1279-1281.
96. Porte D Jr., Kahn SE: **Hyperproinsulinemia and amyloid in NIDDM. Clues to etiology of islet beta-cell dysfunction?** *Diabetes* 1989, **38(11):**1333-1336.
97. Salmona M, Malesani P, De Gioia L, Gorla S, Bruschi M, Molinari A, Della Vedova F, Pedrotti B, Marrari MA, Awan T, Bugiani O, Forloni G, Tagliavini F: **Molecular determinants of the physicochemical properties of a critical prion protein region comprising residues 106-126.** *Biochem J* 1999, **342 (Pt 1):**207-214.
98. Thompson AJ, Barnham KJ, Norton RS, Barrow CJ: **The Val-210-Ile pathogenic Creutzfeldt-Jakob disease mutation increases both the helical and aggregation propensities of a sequence corresponding to helix-3 of PrP(C).** *Biochim Biophys Acta* 2001, **1544(1-2):**242-254.

Recent Structural and Computational Insights into Conformational Diseases

Xavier Fernàndez-Busquets[1], Natalia S. de Groot[2], Daniel Fernandez[2] and Salvador Ventura*[,2,3]

[1]Biomolecular Interactions Team, Nanobioengineering Group, Institute for Bioengineering of Catalonia, and Nanoscience and Nanotechnology Institute, University of Barcelona (Spain); [2]Departament de Bioquímica i Biologia Molecular, Facultat de Biociències and [3]Institut de Biotecnologia i Biomedicina, Universitat Autònoma de Barcelona, E-08193 Bellaterra (Spain)

Abstract: Protein aggregation correlates with the development of several deleterious human disorders such as Alzheimer's disease, Parkinson's disease, prion-associated transmissible spongiform encephalopathies and type II diabetes. The polypeptides involved in these disorders may be globular proteins with a defined 3D-structure or natively unfolded proteins in their soluble conformations. In either case, proteins associated with these pathogeneses all aggregate into amyloid fibrils sharing a common structure, in which β-strands of polypeptide chains are perpendicular to the fibril axis. Because of the prominence of amyloid deposits in many of these diseases, much effort has gone into elucidating the structural basis of protein aggregation. A number of recent experimental and theoretical studies have significantly increased our understanding of the process. On the one hand, solid-state NMR, X-ray crystallography and single molecule methods have provided us with the first high-resolution 3D structures of amyloids, showing that they exhibit conformational plasticity and are able to adopt different stable tertiary folds. On the other hand, several computational approaches have identified regions prone to aggregation in disease-linked polypeptides, predicted the differential aggregation propensities of their genetic variants and simulated the early, crucial steps in protein self-assembly. This review summarizes these findings and their therapeutic relevance, as by uncovering specific structural or sequential targets they may provide us with a means to tackle the debilitating diseases linked to protein aggregation.

Keywords: Conformational diseases, amyloid fibrils, protein aggregation, protein folding, protein structure, Alzheimer’s disease, Parkinson’s disease, prion.

INTRODUCTION

Polypeptide chains are built up of repetitive amino acid units that differ in the chemical nature of their substituents at the α-carbon. Although inside the dense cellular environment amino acid side chains are exposed to manifold interactions, they usually manage to form the specific intramolecular contacts that direct the folding of polypeptides towards the stable, native structure. Such a conformation is necessary for the proteins to perform their biological functions. However, under certain conditions proteins misfold, lose their native structure, and adopt non-native conformations, leading to self-assembly and formation of amyloid protein deposits. The ability to adopt an amyloid-like structure or to undergo fibrillogenesis has emerged as a common, and perhaps fundamental, property of polypeptide chains [1-4]. Although protein aggregates may constitute a stable and secure way to accumulate unwanted proteinaceous material in the cytosol, they can also trigger a cascade of pathological events. Misfolding and protein aggregation are related to more than 30 distinct human conformational disorders [5]. Some examples include the following: Alzheimer’s disease, which is possibly linked to the formation of extracellular plaques and intraneuronal tangles by amyloid β(Aβ) protein and hyperphosphorylated tau protein, respectively; Huntington’s disease, in which long glutamine stretches in huntingtin make it more prone to deposit into intranuclear inclusions and cytoplasmic aggregates; Parkinson’s disease, which is related to α-synuclein aggregation and Lewy body formation; and the spongiform encephalopathies, including Creutzfeldt-Jakob disease, which is associated with human prion protein [6-9].

In many cases, conformational diseases exhibit degenerative pathologies with a major impact on the elderly. In the rapidly ageing developed world, such diseases threaten to lead to a collapse of public health services in the near future. Because of the urgent need to develop measures to prevent and treat conformational diseases, recent years have seen an overwhelmingly vast amount of research dedicated to in depth investigation of events that lead to protein misfolding and to the formation of harmful protein deposits. Structure-focused studies aimed at elucidating the conformation and deposition dynamics of polypeptide molecules in aggregates are likely to yield valuable rewards in the search for specific targets for therapeutic intervention. At the same time, bioinformatic approaches to protein aggregation may reveal detailed information on the mechanisms of aggregation at a molecular level, a challenge for wet lab studies. The methods employed to deduce structural models of amyloid aggregate architecture and the computational tools available to study these protein assemblies are the focus of this review.

NEW STRUCTURAL INSIGHTS INTO CONFORMATIONAL DISEASES

Until recently, one of the main obstacles to characterizing the structure of amyloid fibrils at the molecular level was that they are neither crystalline nor small enough to be studied by solution NMR spectroscopy [10]. However, in the last few years, high-resolution structures of amyloid fibrils have been obtained through major advances in solid-state nuclear magnetic resonance [11-13] and through the use of nano- and microcrystals of small amyloidogenic peptides that can be subjected to single crystal X-ray diffraction analysis [14-16]. A third major technique by which the structure of amyloid fibrils not in solution might be elucidated is transmission electron microscopy (TEM) [17]. Finally, single-molecule methods have recently emerged as a promising new approach for studying amyloid structure [18].

Solid-State Nuclear Magnetic Resonance (SSNMR)

SSNMR is a spectroscopic method ideally suited to study the structure and dynamics of densely packed molecular assemblies at the atomic level [19]. This technique was developed specifically for structural studies of molecular systems in noncrystalline solid and solid-like states such as phospholipid bilayers or precipitated protein aggregates, and it has consolidated as a principal methodology for amyloid fibril analysis (as reviews see [20-22]). SSNMR allows for high-resolution calculation of the distances between ^{13}C labels placed selectively in amyloid-forming peptides [17]. These data are accurate to ca. 0.2 Å for distances less than about 6 Å, although this precision is not maintained for greater distances. Amyloid fibrils are a good system for SSNMR

*Address correspondence to this author at the Departament de Bioquímica i Biologia Molecular, Facultat de Biociències, Universitat Autònoma de Barcelona, E-08193 Bellaterra, Spain; Tel: +34 93 586 81 47; Fax: +34 93 581 12 64; E-mail: salvador.ventura@uab.es

0929-8673/08 $55.00+.00

investigations because (i) although they are noncrystalline, amyloid fibrils do have well-ordered molecular structures and therefore yield SSNMR data that are of high quality and easy to interpret; (ii) they can be prepared with selective or uniform isotopic labeling in the milligram amounts required for most SSNMR measurements; (iii) the fibrils can be obtained in high concentrations, leading to good signal-to-noise ratios; and (iv) the structural information resulting from SSNMR measurements is arguably more direct and specific than information derived from other techniques [21].

The first SSNMR studies of amyloid fibrils were focused on the structure of Aβ fibrils associated with Alzheimer's disease (AD) pathology. In a pioneering study of $A\beta_{34-42}$ fibrils, a structural model of antiparallel β-sheets with an alternating hydrogen bond registry was proposed [23]. Later, analysis of $A\beta_{10-35}$ fibrils [24] revealed for the first time the parallel, in-register β-sheet organization that has since been found in other systems and is now widely believed to be the most common (but not universal) β-sheet organization in amyloid fibrils [25-29], especially those formed by relatively long polypeptide chains. More recently, Tycko and coworkers have undertaken SSNMR studies of fibrils formed by the full-length β-amyloid peptides $A\beta_{1-40}$ [11,30-34] and $A\beta_{1-42}$ [25], and by the fragments $A\beta_{16-22}$ [35,36] and $A\beta_{11-25}$ [36,37]. A model has been put forward for the structure of $A\beta_{1-40}$ protofilaments [21] (Fig. **1**), where each $A\beta_{1-40}$ molecule contributes a pair of β-strands, spanning approximately residues 12-24 and 30-40, to the core region of the fibrils. These strands, connected by the loop residues 25-29, are not part of the same β-sheet, but participate in the formation of two distinct parallel and in-register β-sheets within the same protofilament. In $A\beta_{1-42}$, residues 1-17 are disordered, whereas residues 18-42 form a β-strand-turn-β-strand motif that contains two intermolecular β-strands formed by residues 18-26 (β_1) and 31-42 (β_2) [38].

SSNMR studies have also contributed to the development of a model for the core region of fibrils formed by the prion protein HET-s [13]. In the proposed structure, each molecule contributes four β-strands, with strands one and three forming the same parallel β-sheet and strands two and four forming another parallel β-sheet approximately 10 Å away. This is also the distance separating each of the four β-sheets that form $A\beta_{1-40}$ protofibrils. Additional SSNMR studies that confirm this cross-β-sheet structure of amyloidogenic proteins include work on the transthyretin (TTR) peptide $TTR_{105-115}$ [12,39], the *de novo* designed peptide ccβ [40] and synthetic peptides representing residues 89-143 of mammalian prion protein [41].

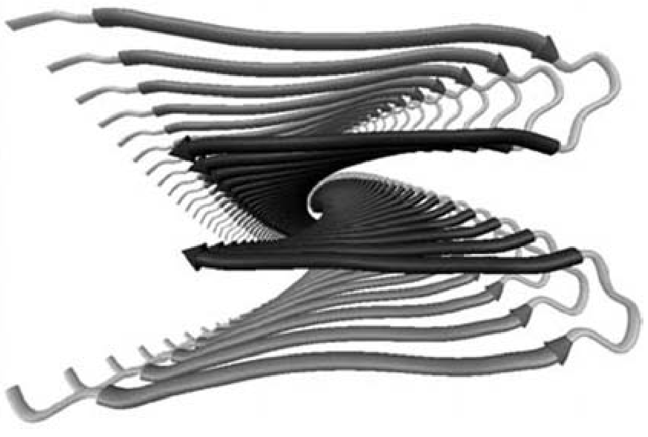


Fig. (1). Ribbon representation of a structural model for the Aβ1-40 protofilament, based on SSNMR and TEM data. In the protofilament, viewed along its long axis, the β-sheets in the cross-β motif are parallel and in-register and are formed by two β-strand segments from each Aβ1-40 molecule (medium and darkest grey segments) that are separated by a loop (lightest grey segment). Two molecular layers form a four-layered structure with a predominantly hydrophobic core [21].

In the neuropathology of AD, Aβ may exert its neurotoxic activity through interactions with the cell membrane [42-44]. Therefore, the study of Aβ-membrane interactions at the molecular level is a key approach towards understanding the pathology of amyloidogenic proteins. Steady-state interactions of $A\beta_{29-42}$ with neutral lipid bilayers have been investigated by SSNMR [45]. Results from these experiments and molecular modeling support the conclusion that when Aβ is inserted in the membrane, its steady-state structure is an oligomeric association of β-sheet peptides located at the bilayer interface. Furthermore, these peptides preferentially recruit phosphatidylethanolamine, suggesting that the inaccessibility of this crucial lipid could be a feature of the cellular disorders induced in AD [22].

X-ray Crystallography

Recently, short amyloid-forming peptides, which exhibit key characteristics of amyloid fibrils, have been successfully induced to form 3D crystals, allowing for the high-resolution elucidation of the packing structure of the peptides [15,16]. The seven-residue peptide GNNQQNY and its relative NNQQNY, both derived from the Sup35 prion protein of *Saccharomyces cerevisiae*, form elongated microcrystals that have made X-ray diffraction studies possible [15]. Atomic structure determination of the cross-β spine revealed that it is a double β-sheet, with each sheet formed from parallel segments stacked in-register. In these crystals, side chains protruding from the two sheets form a tightly self-complementing steric zipper, stabilizing bonding between the sheets. For the longer peptide, the two sheets interact with each other through the side chains of Asn2, Gln4, and Asn6, and these interactions are so tight that water is excluded from the region between them. Within each sheet every segment is bound to its two neighboring segments through stacks of both backbone and side-chain hydrogen bonds. Similarly, fibrous crystals were grown from a 12-mer peptide containing two KFFE motifs separated by an AAAK motif, and they yielded high-resolution X-ray and electron diffraction data [16]. In this case, the peptide associated to form antiparallel β-sheets. However, as in the case of GNNQQNY, these sheets were found to be zipped together via a staggered arrangement of contacts between the side-chains that also excluded water.

Success in the formation of crystals from full-length amyloidogenic proteins is allowing the field of X-ray crystallography to solve a wide variety of structures. X-ray crystallography was used to elucidate the 3D structures of two important TTR variants: TTR Y78F, an amyloidogenic protein, and TTR R104H, a protein associated with a protective effect over the amyloidogenic V30M mutation [46]. While the former structure strongly suggests a relevant role for an α-helix in the overall stability of TTR, the latter suggests that N-terminal stabilization might be the key determinant of its protective effects. Inspection of a limited number of crystal structures of β2 microglobulin (β2m) as an isolated chain separated from the major histocompatibility complex I heavy chain revealed that the corresponding 3D structure is based on an antiparallel β-barrel fold, with an immunoglobulin (Ig) domain topology [47]. The structural bases of amyloidogenic potential in β2m can be related to local unfolding, to the tendency to aggregate laterally through non-compensated β-strands, and partially to its trend towards N-terminal proteolytic degradation [47].

Electron Microscopy

Electrons interact with atoms much more strongly than X-rays or neutrons, thus allowing the observation of individual molecules. The wavelength of the electron beam is much smaller than that of visible light, allowing magnifications ca. 500 times higher than with light microscopes. TEM examination of a substantial number of amyloidogenic proteins has provided images of amyloid fibrils as long, unbranching filaments 6-12 nm across, characterized by the lateral association of protofilaments, and exhibiting a clear helical twist [48-51]. As a drawback of TEM, the intense flow of electrons required to obtain good contrast leads to substantial radiation and sample damage. Staining improves the contrast but it often leads to loss in resolution of internal fibril structure, which adds to the possible distortion induced by dehydration. Cryo-TEM represents an alternative strategy in which the sample is frozen in liquid nitrogen or helium in order to reduce the magnitude of damage from ionization. The proteins are maintained in a hydrated state and low-dose TEM is used to minimize radiation damage. Cryo-TEM does not necessarily give higher resolution than standard TEM, but it provides more reliable structural information, which can be extracted from the micrographs through either direct visualization or 3D reconstruction of the fibril [52]. The protofilament structure of an SH3 peptide was successfully studied by cryo-TEM [53], revealing an elliptical cross-section formed by four protofilaments. Studies performed on lysozyme and apolipoprotein A fibrils indicated varying protofilament arrangements within the same sample [54]. High-resolution cryo-TEM images of $A\beta_{11-25}$ protofilaments revealed striations running across the filament [17], that likely corresponded to individual β strands within a single β-sheet/protofilament, in agreement with Fourier-transform infrared spectroscopy data showing specific bonding patterns consistent with β-sheet structure. A strong reflection at 4.7 Å, which is the hydrogen bond spacing of β strands, was also observed in Fourier transforms of cryo-TEM images of full-length islet amyloid polypeptide (IAPP) fibrils [55], supporting the view that this is a common feature of amyloid fibrils.

Cryo-TEM data has been complemented by information derived from scanning TEM (STEM), which allows the quantitative determination of mass-per-unit length of a fibril by comparison with a standard, such as the tobacco mosaic virus [56]. If the specimen is thin, the image intensity is directly proportional to the mass of the irradiated region. STEM measurements of $A\beta_{1-40}$ protofilaments were consistent with the peptide being folded on itself and stacked to form two β-sheets in contact with each other through their side-chains [11]. The cross-β structure deduced for IAPP protofilaments from X-ray diffraction data [57] was also confirmed by diameter and mass-per-unit length measurements determined using TEM and STEM [55,58].

Image reconstruction from electron micrographs has been used to examine the structure of TTR amyloid fibrils [59]. Averaged cross-sections of 200 different fibrils produced a detailed view of the substructure, revealing a fibril diameter of ca. 130 Å. Cross-sections of the fibrils exhibit 4-fold symmetry with four proto-filaments, each measuring 40 to 50 Å across, arranged around a central hollow core. 3D maps of SH3 fibrils based on cryo-TEM data have also revealed a hollow core with a maximum diameter of 5 nm [53]. Other 3D structures reconstructed from TEM or cryo-TEM images include those of fibrils formed by mammalian prion protein [60] and insulin [61].

Whereas SSNMR, X-ray crystallography, and TEM have provided valuable high-resolution data on the structure of amyloid fibrils, they have important limitations. Some of the most fundamental characteristics of amyloid fibrils that are still largely unknown include their dynamics and kinetics of growth and disassembly, their stiffness, their mechanical response to compression, tension, or pulling and their resistance to breakage, among others. These properties are being increasingly acknowledged to be important to understanding the *in vivo* roles of amyloid deposits [62-64]. Study of the conformational plasticity of amyloids has required new experimental approaches that materialized with the development of nanometric methods [18]. The advantages of most nanometric techniques arise from their ability to observe and/or manipulate single molecules in a liquid environment.

Single-Molecule Studies

Single-molecule approaches have the capacity to provide previously unattainable data on elementary biological processes. Until recently, most experimental techniques derived data from molecular ensembles: examples include bands in electrophoresis and in Western, Southern, and Northern blots; ordered structures in NMR and crystallographic studies; absorbance, fluorescence, or diffraction in solution studies; and optical microscopy. Although these methods have obviously generated and will continue providing essential structural and functional information, the data obtained only represent the mean values of large numbers of molecules. This “molecular sociology” is tremendously informative for the understanding of biomolecular processes. However, just as human sociology is adequate to study groups of people, other disciplines like anatomy are much more appropriate for the study of the individual. Thus, single-molecule techniques can be considered to be the study of “molecular anatomy” in the sense that they explore and manipulate individual molecules. Clearly, single-molecule techniques can tap an unfathomable ocean of new information that can not be obtained with “multiple-molecule” approaches. In a recent survey of literature employing single-molecule studies [65], a search on PubMed (www.pubmed.gov) revealed exponential growth over the last two decades. Conventional multi-molecule tools can only provide an averaged picture of a system under study, where much of the subtle or short-lived information is lost. Emerging nanotools might complement this limitation by opening novel paths for the development of early diagnostic and therapeutic approaches [66].

Application of Nanometric Methods to the Study of Self-Aggregating Proteins

In order to fully understand the process leading to fibril deposition it is paramount to have access to methods that allow the study of the growth and/or disassembly of individual amyloid fibrils. Such knowledge will be essential to the design of novel therapeutic methods for the removal of amyloid aggregates. Among the plethora of an ever-growing array of single-molecule approaches, we will focus our attention on those that are being currently employed to explore amyloid structure and dynamics. These include (i) scanning probe microscopies such as scanning tunneling microscopy (STM) and atomic force microscopy (AFM), the latter of which is used both for imaging and for single molecule force spectroscopy (SMFS); (ii) optical methods based on the properties of the evanescent wave, such as total internal reflection fluorescence microscopy (TIRFM) and near-field scanning optical microscopy (NSOM); and (iii) fluorescence correlation spectroscopy (FCS).

Scanning Probe Microscopy (SPM) Methods

SPM methods are a group of techniques where a surface is imaged at high (and in some cases atomic) resolution by rastering an atomically sharp tip in close, but not direct, contact with the surface [67]. The strength of the interaction between the tip and the surface and the relative position of the tip is measured to produce an image of interaction strength as a function of position, which, depending on the particular technique, represents surface topography or chemistry. SPM allows investigation of structural as well as functional properties of native biomolecules in liquid and physiological environments by a unique combination of subnanometer spatial resolution, millisecond temporal definition and piconewton force sensitivity [68-70]. This method relies on the highest possible precision for the movement of the cantilever holding the tip, which is achieved by the use of piezoelectric transducer elements and whose accuracy is well below 1 Å (smaller than one single atom). Most commonly, the cantilever deflections are monitored via a laser beam, where the reflected laser spot is converted on a position sensitive photodetector into an electric signal.

In scanning tunneling microscopy (STM) a voltage is applied between the cantilever tip and a conductive or semiconductive sample. Under these conditions, electrons can flow between the probe and the surface, generating an electric "tunneling" current whose intensity is inversely proportional to the distance that it has to span. A 3D map is generated by fixing a constant value for the current between the probe and the specimen and simultaneously recording the vertical displacement of the cantilever during the scan. Despite the subnanometer resolution that can be obtained by STM, the application of this technique to biological molecules is limited by the requirement that the imaged sample be conductive, although significant progress has been made in imaging Aβ fibrils by STM [71-73].

In atomic force microscopy (AFM), forces between the atoms of the scanning tip and those of the surface-immobilized molecules under examination induce the vertical displacement of the cantilever. The topographic image of the scanned sample is generated by monitoring these signals with a spatial resolution that is equal to or even higher than that obtained for the same sample with the electron microscope [74], but with the benefit that SPM images can be obtained with staining-free protocols on functional biomolecules in physiological solutions. The relative large radius of the AFM tip apex respective to the size of the molecules being scanned requires mathematical deconvolution of raw data in order to calculate the real horizontal dimensions. On the other hand, although STM can provide highly precise distance measurements in the *x-y* plane, it is much less accurate than AFM for height measurements in the *z* axis. AFM has been widely used to monitor the assembly of numerous amyloid-forming peptides and proteins [75], including Aβ, Ig light chain, α-synuclein, β2m, IAPP, insulin, the B1 domain of protein G, TTR, lysozyme, and the 90-residue Src homology (SH3) domain of the α subunit of bovine phosphatidylinositol-3'-kinase [49,50,76-80].

AFM imaging in real time (also termed on-line or time-lapse AFM) allows for the identification and tracking of prefibrillar structures and the continuous monitoring of individual molecular assemblies in a liquid environment with temperature and buffer control. Such studies have provided valuable data on elongation rates, directionality of growth, changes in morphology for individual fibrils, and the assembly process of higher order polymorphic species. One application of time-lapse AFM in this area has been the direct visualization of amyloid fibril growth *in vitro* [81]. This experimental approach holds promise for the future testing of potential therapeutic drugs, for example, by directly visualizing at which level of fibril assembly (nucleation, elongation, branching, or lateral association of protofibrils) a given active compound will interfere.

Perhaps the main potential of AFM is its ability to be used as a tool for the manipulation of biomolecules. Direct manipulation of individual amyloid fibrils or monomers can provide essential information on aspects such as fibril brittleness or elasticity, or on the adhesion forces that govern the self-aggregation process of individual polypeptide molecules.

Single Molecule Force Spectroscopy

During the last decade SMFS has developed into a highly sensitive tool for investigation of the interaction of single biomolecules [82]. Most SMFS experiments use either optical tweezers or AFM to measure piconewton dissociation forces of single ligand-receptor complexes. The molecular binding partners are attached to the micro- or nanoscale force sensor and a sample holder, respectively, by covalent chemistry. When both moieties are brought into close contact, a specific bond between the individual molecules can form. By increasing the distance between the two surfaces again, the molecular bond is loaded under an external force until it finally breaks, yielding the molecular dissociation force. By systematically varying the externally applied load and monitoring the mechanistic elasticity of the complex, information about the kinetic reaction rates, the mean lifetime, the equilibrium rate of dissociation, dissociation length and the energy landscape of the interaction can be derived [83,84]. AFM has the sensitivity to measure forces comparable with that of a single hydrogen bond [85], a magnitude far below the force exerted in most enzyme-substrate interactions.

SMFS-based manipulation methods have been applied to explore the mechanics and structural dynamics of amyloid fibrils [86]. In mechanically manipulated individual amyloid fibrils formed from either $A\beta_{1-40}$ or $A\beta_{25-35}$, β-sheets behave as elastic structures that can be "unzipped" from the fibril with different amounts of constant force. Unzipping was fully reversible across a wide range of stretch rates provided that coupling, via the β-sheet, between bound and dissociated states was maintained. These data suggested that the rapid, cooperative zipping together of β-sheets could be an important mechanism behind the self-assembly of amyloid fibrils. The use of appropriate surface chemistry enabled the anchoring of Aβ through the N-terminal ends, allowing the measurement of the rupture of Aβ-Aβ contacts at single molecule level [87]. The rupture of these interactions was accompanied by the extension of the peptide chain detected by a characteristic elasto-mechanical component of the force-distance curves. SMFS has also been used to study the effect of pH on the interactions and misfolding of α-synuclein, Aβ, and lysozyme [88], showing that the attractive force between homologous protein molecules is minimal at physiological pH and increases dramatically in an acidic milieu.

The severing of amyloid fibrils generates seeds for new fibril formation, and thus represents a key determinant of their physiological impact [64]. In an elegant approach, a detailed mechanical characterization of individual insulin amyloid fibrils revealed that they have a strength comparable to that of steel and a mechanical stiffness comparable to that of silk [89], indicating that amyloid fibrils possess properties that make them potentially useful materials for biotechnological applications. In the study, insulin fibrils were deposited on a silicon surface that had been nanopatterned with grooves. Then, a fibril spanning a groove was selected by AFM imaging and force-distance curves were acquired from different spots along the suspended fibril by applying a given load with the AFM cantilever. This analysis revealed that the forces required to mechanically fracture amyloid fibrils range from 300 to 500 piconewtons (pN) [89], which are values on the same order of magnitude as the forces required to unfold individual protein domains [90] or to break carbohydrate-carbohydrate interactions in cell-to-cell proteoglycan-mediated adhesion [91]. This suggests that the interactions within the amyloid core, which include hydrogen bonding, van der Waals forces, and electrostatic forces, are of similar origin and nature as those responsible for the folding of native structural motifs in proteins. The high level of stability of amyloid aggregates is an essential factor underlying their involvement in a range of clinical disorders [92]. The high strength and mechanical stiffness of amyloid fibrils makes them very resistant to degradation by the endogenous mechanisms of living organisms, thus leading to an accumulation of protein aggregates.

Optical Microscopy Methods

The ~250 nm resolution limit of conventional optical microscopy was the primary factor behind the development of higher-resolution electron microscopy and scanning probe techniques. These and related microscopic methods have enabled phenomenal gains in resolution, up to the level of visualizing individual atoms [93]. However, the requirement for highly purified molecular preparations limits the application of these high-resolution methods in many biological investigations, especially those measuring *in vivo* dynamics.

A first approximation to obtaining nanometer resolution in live-cell optical fluorescence microscopy was the development of TIRFM. In this method, when a beam of light hits the interface between two media of different refractive index at above a critical angle, all of the light is reflected but some of the incident energy enters the second medium forming an electromagnetic field that oscillates with the same frequency as the incident light, generating the so-called evanescent wave. For a laser light at 455 nm, this evanescent field penetrates only for ca. 150 nm into the sample [94], although this distance is also a function of the refraction indexes and the angle of incidence beyond the critical angle. As a result, fluorescent molecules are excited within a very small volume near the interface and there is a low background of out-of-focus fluorescence compared to epifluorescence, from which TIRFM derives its potential for detecting single molecules. Finally, the excitation light, which can also contribute to background noise as it is usually much more intense than the emitted fluorescence, is cleanly removed from the TIRFM image, as any that is not absorbed gets carried away in the reflected beam [95].

Individual amyloid fibril growth can be visualized in real time by following the incorporation of protein monomers or oligomers into preformed seeding fibrils or protofibrils. Polymerization of the amyloidogenic yeast prion protein Sup35 has been studied using this approach [63,96]. Sup35 fibers labeled with a red fluorescent dye were deposited on a slide and treated with a solution of monomeric Sup35 labeled with a green fluorescent dye. Fibril growth observed both by epifluorescence [96] and by TIRFM [63] occurred mainly unidirectionally through direct monomer addition, in the absence of observable intermediates. By monitoring Thioflavin T (ThT) fluorescence with TIRFM, the growth of amyloid fibrils formed by β2m could be followed without the need for covalent fluorescent labeling [97]. The results obtained showed that the extension of β2m fibrils was mostly unidirectional. Since ThT binding is common to all amyloid fibrils, this method will likely have general applicability. ThT fluorescence measurements have also been used to study the dynamics of fibrillogenesis for an octapeptide derived from the C terminus of human medin and for $A\beta_{1-40}$ [97].

For live cell imaging, TIRFM illuminates only the basal membrane proximal to the microscope slide, imposing a constriction on the experimental design. Yet, TIRFM can provide valuable dynamic information on the interaction of living cell membranes with different oligomeric or fibrillar amyloid species with which the slides can be previously functionalized. There are evidences indicating that the appearance of insoluble fibrillar structures enriched in β-sheets is facilitated by diverse environmental factors [98], biological membranes among them. Aβ is cleaved from its precursor protein in the membrane interface, and its cytotoxic effect is likely related to the amyloid-lipid interaction [22]. TEM and AFM studies performed with artificial membranes have shown that, upon interaction with the membrane lipids, Aβ in fibrillar form reverts to soluble globular peptide oligomers that associate into disordered domains [99]. A detailed description of the role in amyloidogenesis of membranes and membrane constituents falls outside the scope of this work, and the interested reader can find that information in a number of recent detailed reviews [22,100,101].

Near-Field Scanning Optical Microscopy

In NSOM, as in the case of AFM, a sharp probe scans the sample surface, but in addition to topography, NSOM also generates optical images (for a review see [102]). The most generally applied near-field optical probe consists of a small aperture, typically 20-120 nm in diameter (i.e., much smaller than the wavelength of the excitation light), at the end of a metal-coated tapered optical fiber. In fluorescence mode, it serves as a constriction that funnels an incident light wave to dimensions that are substantially below the diffraction limit, resulting in a light source that has the size of the aperture. However, in contrast to common light sources such as lightbulbs and lasers, the light emitted by the probe is predominantly composed of evanescent waves rather than propagating waves. As described above, the intensity of the evanescent light decays exponentially and to insignificant levels ~100 nm from the aperture, and thus the probe can only excite fluorophores that reside within a layer of <100 nm from the probe, in what is termed the near-field region. Sample fluorescence can subsequently be collected by conventional optics and transformed into an optical image of the sample surface in which the resolution is now primarily dictated by the aperture dimensions rather than by the wavelength of the light. An electronic feedback system keeping the probe-sample distance during scanning at less than 10 nm is used, as in AFM, to generate a topographic map of the sample surface. Unique to NSOM is the fact that a corresponding fluorescence map is simultaneously generated. Although NSOM has not been widely applied yet to the study of amyloidogenesis, it is foreseeable that in the near future this technique will provide valuable data on the dynamics of the interaction of different amyloid species with membranes.

Fluorescence Correlation Spectroscopy

FCS is based on the fluctuation of light emitted by fluorescently labelled molecules crossing a small laser spot and detected with confocal optics [103,104]. Fluorescence fluctuations are the result of molecular diffusion, chemical reactions, and physical processes of a few fluorescent molecules in an optically restricted sub-micron observation volume (~ 1 fL $= 10^{-15}$ L), that can be studied with a temporal resolution in the range of 1 ms to >10 s. The elegance of FCS lies in its ability to extract a wealth of molecular and environmental data from a weak signal that is comparable with background noise. This information is obtained by using correlation analysis of the fluorescence fluctuations of very small samples of molecules at nanomolar concentrations. Recent technological advances have enhanced the number of biological and chemical applications of FCS, such as study of binding interactions between biomolecules, sparse molecule detection, intramolecular protein dynamics, and diffusion in the membranes of living cells.

The first FCS data on the kinetics of amyloid formation reported on the cooperativity of Aβ polymerization [105], showing that the formation of very large aggregates preceded the formation of fibrils. FCS has been recently applied to the study of the interaction of Aβ with cell membranes [106]. Here, rhodamine labelled Aβ showed different diffusion times that likely corresponded to Aβ either unbound or complexed to a target molecule in the cell membrane, thus adding to the hypothesis that Aβ affects neurons through its binding to a receptor [107]. It appears that Aβ binding to some membrane proteins may be protective for the cell, e.g. by mediation of Aβ internalization and degradation [108-110]. On the other hand, Aβ binding to certain plasma membrane receptors can damage the cell by promotion of tau protein phosphorylation [111], generation of oxidative stress and stimulation of macrophages [112-115], blocking of protein function [116,117], or induction of apoptosis [118]. Single Aβ aggregates could also be detected in the cerebrospinal fluid of AD patients by FCS [119], suggesting that this technique could enable easy *in vivo* detection of cerebral amyloid and might hold potential value for enhancing routine diagnosis. FCS has also been applied to study the oligomerization of other amyloidogenic polypeptides such as the polyQ stretch within disease-causing proteins [120]. In both cases, FCS-derived data on the dynamics of individual fluorescent molecules in solution provided valuable insights into the respective polymerization processes.

Amyloid Fibril Structure Deduced from SSNMR, X-ray, and TEM Data

The similarities between amyloid fibrils derived from very different peptides and proteins suggest that their core structures have similar features primarily dictated by the intrinsic conformational preferences of polypeptide chains [10]. However, the specific nature of the side-chain packing, including such characteristics as the alignment of adjacent strands and the separation of the sheets, provides an explanation for the occurrence of variations in the structural details of different types of fibrils. Putting together the above data derived from SSNMR, X-ray crystallography, and TEM, a picture of the common properties of amyloid fibrils emerges that includes, as a universal element, the cross-β X-ray diffraction pattern. This pattern consists of an X-ray reflection at ~4.8 Å resolution along the fibril axis and another X-ray reflection at ~10-12 Å resolution perpendicular to the fibril [121], indicating that the fibrils contain β-sheets parallel to the fibril axis with their extended protein strands perpendicular to the axis. Amyloid structures can be built of parallel or antiparallel sheets, that in turn may or may not be in register [14]. Unfortunately, if the strands are out of register rather than being antiparallel or if the sheet is composed of a mixture of parallel and antiparallel strands, misleading results may be generated [17]. This might explain the wide variety in reported results. To date, antiparallel β-sheet structures have been identified only in fibrils formed by relatively short peptides containing only one β-strand segment, indicating that studies of model peptides can not generally be used to infer structural properties of full-length amyloid-forming sequences [21]. The more fundamental feature, which is common to all amyloid and amyloid-like fibrils, appears to be the dry steric zipper motif present in the structure of the cross-β spine [122].

Other Models Besides the Cross-β Structure

Some proposed models are composed largely of structural motifs that are present in the monomeric, native form of the protein [14]. Evidence for retention of native-like structure has been demonstrated for fibrils of Ure2p [123,124], TTR [125], β2m [126], and RNAse A [122]. Thus, at least some amyloid-forming proteins might not have two distinct, stable structures (i.e. a native state and an amyloid state). To the extent that native-like structures are found in general as part of amyloid-like fibrils, the changes in structure would be mainly confined to the segments forming the steric zipper.

Other alternative models center around a β-helical or nanotube structure that has been suggested as a possible generic architecture for the amyloid fibril (reviewed in [14]). In these models, one or more extended β-sheets wrap around a hollow core in a helical manner. Antiparallel β-helix models have been inferred for amyloid fibrils formed from the peptide KLKLKLELELELG [127], and from TTR, Aβ, and Ig light chain [128]. Further evidence for β-helical models has been provided for a substantial number of different polypeptides, through several different experimental approaches [129-133].

TEM structural analysis of Aβ_{1-40} protofilaments showed a mixture of straight and twisted fibers [134]. The average width of both types was ~70 Å, and the helical pitch of the latter was ~460 Å. Cross sections of embedded samples showed a ~60 Å-wide tubular species. X-ray diffraction from these samples indicated the presence of the cross-β fiber pattern characterized by a strong meridional reflection at 4.7 Å and a broad equatorial reflection at 8.9 Å. Modeling studies suggested that tilted arrays of β-strands constitute tubular, 30 Å-diameter protofilaments, and that three to five of these protofilaments constitute the Aβ fibril. This type of multimeric array of protofilaments organized as a tubular fibril resembles that formed by the shorter Aβ_{6-25}, Aβ_{11-28}, and Aβ_{1-28} [135-137], which suggests a common structural motif in Aβ fibril organization.

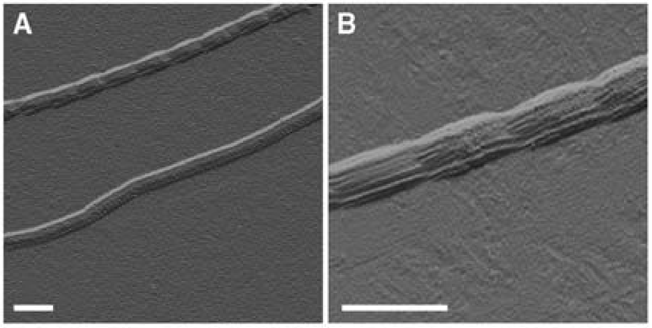


Fig. (2). Amplitude AFM images of nodular and smooth Aβ1-42 fibrils. A) Nodular and smooth fibrils. B) Nodular fibril from panel A scanned at higher resolution. Bars: 100 nm [50].

The accumulated evidence described above in favor of the existence of both protofilament helical structures and tubular fibril forms suggests that both species are different states within a dynamic fibrillogenesis pathway. Such a dynamic process has been proposed to have as an end-point the so-called smooth fibrils [50], which would have as immediate precursors a nodular type of fibril (Fig. **2**). Nodular fibrils are constituted of ~100 nm-long segments that are defined by internal helical structures with a pitch of also ~100 nm. These helical structures have dimensions consistent with the protofilament helix model shown in (Fig. **1**).

Fig. (**3A**) shows an AFM image of partially assembled/disassembled fibrils, where hints of the fibril nodules are observable by the presence of an apparently softer material that can be significantly pushed aside when applying greater forces with the AFM tip (Fig. **3B**). This manipulation of fibrils revealed the existence of an underlying protofilament helix with a period that coincides precisely with the length of nodules in the fibril being imaged. Furthermore, AFM and TEM images show fibrils that have been fractured in sections (Fig. **3C**) with a mean fragment length of 107.3 ± 29.0 nm, which is close to the measured periodicities of fibril nodules (93.5 ± 21.0 nm) and of the helical repeat of intertwined protofilaments (92.5 ± 20.3 nm). In these segmented fibrils the two clearly discerned protofilaments run parallel and do not appear to twist; also, the protofilament sections strongly resemble protofibrils ~100 nm long. Although such fibril fragmentation might be due to the sample manipulation, they demonstrate the existence of a structural weakness related to the joining points between the constituent ~100-nm protofibril subunits.

Taken together, these data are consistent with the existence of a ~100 nm-long motif that is a key intermediate in the fibril assembly process. Such an intermediate may be found not only in Aβ, but possibly in many amyloid fibrils, as suggested by the existence of structural units of similar length formed by other amyloidogenic proteins and peptides such as SH3 [138], the prion protein [60,139], and β2m [79]. A structural organization similar to that observed for Aβ_{1-42} in protofilament helices and fibril nodules has been described for other types of amyloid fibrils produced *in vitro* from very different polypeptide sequences such as lysozyme and SH3 [49]. Although Aβ, lysozyme, SH3, and a number of other amyloidogenic proteins and peptides [140] are not sequence-related, they form twisted protofilaments and nodular fibrils remarkably similar in periodicity, dimensions, and number of constituent subunits. This is in agreement with the current view that, despite the different nature of precursor proteins and peptides,

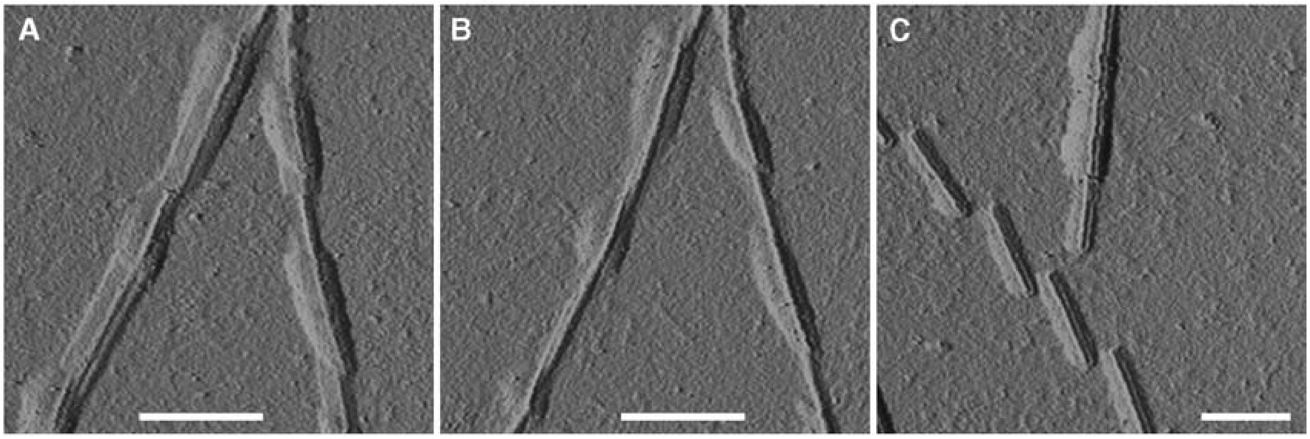


Fig. (3). AFM manipulation of Aβ1-42 fibrils and visualization of segmented fibrils. A, B) AFM images taken on HOPG of the same fibrils scanned with (A) low and (B) high amplitude. Higher amplitude increases the force between tip and sample. (C) Amplitude AFM image of a segmented fibril. Images are presented as unprocessed amplitude signal, where increasing brightness indicates greater damping of cantilever oscillation. Bars: 100 nm [50].

amyloid fibrils represent a structural superfamily and share a common protofilament structure [77,121,140].

NEW COMPUTATIONAL INSIGHTS INTO CONFORMATIONAL DISEASES

The increasing knowledge on the sequential and structural constraints in protein aggregation, together with a solid classical background on the determinants of protein folding and stability, has allowed the recent development of a number of theoretical methodologies to predict and model protein aggregation. In this second part of the review we shall try to describe some of the new *in silico* approaches to studying protein deposition. First, we will discuss predictive methods relying on the analysis of the physicochemical and/or structural properties of amino acids in amyloidogenic protein sequences. Then, we will address approaches based on the experimental determination of protein aggregation propensities within amyloidogenic sequence stretches. Emphasis will be given to algorithms published in the last few years. For the description of already classical programs such as those developed by Serrano's, Caflisch's and Dobson's groups, the reader is directed to excellent reviews published elsewhere [141-143]. We will close this section with a review of recent insights on the mechanistic details of protein self-assembly and deposition, as determined by *in silico* simulation methods.

Prediction Methods Based on Sequence-Structure Relationships in Local Protein Regions

The prediction of aggregation-prone regions in polypeptides does not necessarily need to rely on complex assumptions. A recent example is the program SALSA [144]. It calculates the average β-strand propensity score of a peptide window, which the authors name ''β-strand contiguity'', by a very simple treatment of Chou and Fasman's secondary structure preference numbers [145]. Despite employing just a single physicochemical property of amino acids, the authors demonstrate that peaks in the SALSA plots correlate well with the location of amyloid fibril cores in three pathogenic proteins: α-synuclein, $A\beta_{1-40}$ and tau protein. The simplicity of this approach permits fast identification of protein regions with latent β-strand propensity but does not allow the prediction of global polypeptide aggregation or deposition rates because they are influenced by other intrinsic protein properties, such as global charge and hydrophobicity [146].

In contrast to SALSA, Yoon and Welsh [147,148] based their approach on the hypothesis that the propensity of individual amino acids to adopt a particular secondary structure is influenced by their overall tertiary environment, in addition to their simple physicochemical properties. Therefore, it was argued that secondary structure propensities cannot be determined directly from the sequence, and structural information is also necessary. They quantified the influence of tertiary effects on secondary structural preferences by using a simplified approach that counts the number of atom-to-atom tertiary contacts between nonadjacent residues that are spatially close to one another in the native conformation of a given protein. The benchmark used in the study was SCOP20, a collection of protein domains that exhibit <20% sequence identity between any two members. The sequence-structure relationship of any query sequence was systematically evaluated in terms of tertiary contacts by analyzing the secondary structure preferences of similar template sequences in the database. They use this conformation-based approach to detect nonnative (hidden) sequence propensity for amyloid fibril formation. It assumes that segments with high amyloidogenic propensity should display elevated tendency towards β-sheet formation in tightly packed environments (i.e., those with a high number of tertiary contacts). The method correctly assigns high scores to minimal peptide fragments shown experimentally to mediate amyloid fibril formation in Aβ peptide, IAPP, α-synuclein, and human acetylcholinesterase.

Protein structural information allowed Galzitskaya and coworkers [149] to derive a new parameter, mean packing density (number of residues within a given distance from the considered residue), to be incorporated in the prediction of amyloidogenic and intrinsically disordered regions in protein sequences. The mean packing density of each amino acid was derived from the analysis of a database of protein structures, looking at the number of residues close to any given non-covalently bound neighbour. The observed mean packing density was found to be maximal for the three aromatic residues, Tyr, Phe, and Trp, which have been shown to be relevant for aggregation in different polypeptide systems. The authors realized that protein regions possessing strong packing density correlate with aggregating sequences, which presumably intersect with amyloid promoting regions in proteins. In contrast, regions with weak packing density correspond in many cases to disordered regions of proteins. For any query sequence, the expected packing densities are averaged over a sliding window, and a packing density profile is produced. A region is predicted as amyloidogenic if the expected packing density is above a certain, calibrated, threshold. For eight out of twelve examined disease-related proteins and peptides, the predictions were consistent with experimentally tested amyloidogenic regions.

As discussed above, a breakthrough advance in the amyloid field is the resolution of the atomic structure of a common amyloid cross-β spine involving the formation of steric zippers [15]. This information has opened new avenues of research and paved the way to novel predictions that make use of precise structural information. The central idea in these new structure-based methods is that the formation of the cross-β spine in a fibril might be achieved by a segment of the protein, independent of its location within a particular secondary-structure element [150].

Eisenberg, Baker and co-workers used a structure-based approach for the prediction of fibril formation starting from the crystal structure of the fibril-forming peptide NNQQNY from Sup35 [151]. They assumed that a six-residue sequence stretch was sufficient to drive polypeptide amyloid formation. To identify those segments that might be capable of nucleating fibrillogenesis they used 3D profiling. The side chains in the cross-β spine of NNQQNY were mutated *in silico* to those of the sequence of interest and the energetic fit of these variants to the template ensemble was evaluated. Hexapeptide segments that fit well into the template were selected using the ROSETTADESIGN algorithm energy function [152]. This function incorporates contributions from apolar interactions, hydrogen bonds, and steric overlaps. Because it includes these various factors, it can identify fibril-forming sequences that would not have been selected on the basis of simple intrinsic properties, such as hydrophobicity or β-strand propensity. Fibril-forming segments that have been experimentally observed for lysozyme, muscle myoglobin, $A\beta_{1\text{-}42}$, and tau were correctly predicted by this method. Interestingly enough, the method identified fibril-forming regions irrespective of their secondary structure context in the native polypeptide. Thus, segments other than native β-strands of lysozyme and muscle myoglobin, which is mainly an α-helical protein, were identified.

Saiki and co-workers [153] described a method for the evaluation of the propensity to amyloidogenicity using a simplified structural model based in two essential assumptions: The first one is that hydrophobic and hydrogen-bonding interactions occur between residues on neighbouring β-strands aligned in an antiparallel orientation. The second assumption is that interacting hydrophobic residues are present at both faces of the protofibril. The latter condition assumes that the line-matching interactions at both surfaces of a β-sheet protofibril are essential for repetitive stacking. To construct a model for the sheet-to-sheet lamination, the authors looked at interactions between β-sheets in globular proteins and selected the antiparallel system because it is found at a relatively high frequency (e.g. in the Greek-key motif). To model the interactions among paired β-strands, they developed a multi-term mathematical expression, which takes into account the hydrophobic interactions of coupled residues and hydrogen bonding interactions, and introduces one term that checks the required coupling of hydrophobic interacting residues at each face of a β-sheet. Twelve polypeptides or protein fragments known to form amyloids were evaluated using a sliding-window technique. The reliability of the method was assessed by experimentally characterizing peptides that according to the predictions have some degree of amyloid formation propensity. Most of these peptides showed fibrillogenic properties. Conversely, peptides predicted to have low amyloid formation propensity were shown to be unable to form amyloids.

Along this line of thought, Trovato and co-workers [154] derived the PASTA algorithm. In this approach, the propensities of two residues to be found within a β-sheet facing one another on neighbouring strands were determined from the analysis of a dataset of globular protein structures. The model assumes that distinct protein molecules involved in fibril formation will adopt the minimum-energy β-pairings in order to better stabilise the cross-β core of the fibril. Two identical protein chains are assumed to associate by means of an ordered pairing of two hydrogen-bonded β-strands of the same length, whereas the remaining parts of the polypeptide chains remain unstructured. All possible pairings are studied by sliding the two strand-forming regions along the corresponding sequences and varying their length and their relative orientations. The two possible orientations, parallel and antiparallel, are considered. In summary, PASTA does a comparison of the energy score of the parallel and antiparallel β-pairings of a sequence stretch with itself. The authors realize that the parallel in-register arrangement provides a natural way of maximizing the number of favourable stacking interactions, lining up hydrophobic and hydrophilic residues in long rows along the fibril axis. PASTA was employed to analyze the sequence of five natively unfolded systems, namely $A\beta_{1\text{–}40}$, α-synuclein, IAPP, the PHF43 fragment from tau, and the HET-s prion domain. Good agreement between the predictions and the experimental information available on these amyloid structures was found for most systems.

Prediction Methods Based on Experimental Aggregation Propensity Measurments

Although, as described above, theoretical approaches based on structural or sequential intrinsic protein properties have achieved a high accuracy in their predictions, extrinsic factors can dramatically modulate polypeptide aggregation propensities *in vitro* and especially *in vivo*. Hence, several groups have focused their efforts on developing algorithms using *in vitro* or cell-based experimentally determined aggregation propensities, employing different amyloidogenic protein systems. Rojas Quijano and co-workers [155] have used the tau protein as a model. Tau is a largely unstructured protein, but the aggregates it forms display all the physical characteristics of amyloids. Several short nucleating regions have been identified experimentally in tau, among them PHF6 ($_{306}$VQIVYK$_{311}$). The goal of the study was to formulate a model in which the chemical properties of amino acid residues substituted at a single site in the PHF6 structure (VQIVXK, where X was individually substituted by the 20 standard amino acids, except cysteine) could be correlated to the propensity of each peptide mutant to form amyloid fibrils. The kinetics, conformation and morphology of the aggregates formed by the 20 different hexapeptides were analyzed *in vitro*. This allowed for the extrapolation of a scale of amino acid aggregation propensities, which was used to successfully estimate the amyloidogenic propensities of sequences within the tau protein capable of amyloid formation.

The *in vivo* aggregation of polypeptides does not necessarily have to correlate with their *in vitro* properties, since polypeptides within the cell encounter a highly complex and crowded intracellular environment where the protein quality control machinery modulates the accumulation of aggregation-prone polypeptide chains by facilitating their folding, masking hydrophobic regions and targeting improperly folded proteins towards degradation pathways [156].

Measuring *in vivo* polypeptide folding and aggregation has traditionally been a challenging task. Recently, several molecular probes capable of indicating the solubility of target proteins within living cells have been engineered. Using these probes, Waldo and co-workers [157] showed that the folding trajectory of a protein of interest fused upstream to the green fluorescent protein (GFP) dictates the fluorescent behavior of the reporter protein in such a way that the emitted fluorescent signal is directly proportional to the amount of correctly folded target protein. We sought to exploit this approach to study a complete set of mutations in one of the best characterized "hot spots" of aggregation in a disease-linked polypeptide: the central hydrophobic cluster (CHC) of Aβ. We fused the $A\beta_{1\text{-}42}$ protein upstream of the GFP and expressed it, as well as 19 other mutants differing only in a single side chain at the central CHC residue, inside prokaryotic cells. We confirmed that the cellular levels of GFP fluorescence depended exclusively on the

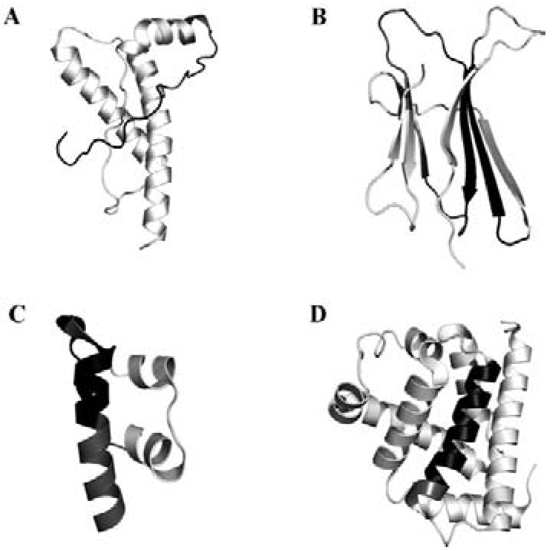


Fig. (4). AGGRESCAN predicted aggregation-prone sequences in amyloidogenic proteins. Structural representation of different proteins showing in black the regions where the AGGRESCAN predictions coincide with the experimental data and in grey the regions where experimentally derived data were not predicted by the algorithm. A) Human prion protein (PDB 1QLX), B) Human β2m (1LDS), C) Human insulin (2OMI), and D) Horse heart myoglobin (1AZI).

in vivo aggregation propensity of the $A\beta_{42}$ variant. The study allowed us to obtain for the first time an individual intrinsic aggregation propensity scale of the 20 common amino acids in an *in vivo* context [158]. These values were further used to develop a simple approach able to accurately predict protein fragments involved in the aggregation of disease-related proteins and the effect of genetic mutations on their deposition propensities [159]. The AGGRESCAN algorithm was implemented into a web server [160] that allows the simultaneous analysis of the aggregation properties of large sets of protein sequences. This analytical ability might be important for protein production in large-scale structural initiatives, for the analysis of the distribution of aggregation-prone regions in complete genomes or for evolution studies, since it is likely that natural protein sequences have evolved in part to code for avoidance of aggregation. The application of AGGRESCAN to the prediction of aggregation-prone regions in globular proteins is illustrated in (Fig. **4**).

Remarkably, a comparative analysis of the aggregation propensities obtained by theoretical calculations with those obtained experimentally [160] shows that, despite the diversity of approaches, there is a striking correlation between *in vitro*, *in vivo* and *in silico* data [160]. Likely, this is because aggregation depends on a combination of characteristics, such as hydrophobicity, charge, secondary structure propensity and packing density, all of which are included in either an implicit or explicit way in most algorithms (Table **1**). As a consequence, the performance of programs like AGGRESCAN, not aimed at the specific identification of short amyloidogenic peptides, but rather of aggregation prone sequences within large natural proteins, compares well with those of pure structural algorithms, when analyzing databases of very short peptides containing both amyloid formers and nonformers [160]. Accordingly, as shown in (Table **1**), most reviewed algorithms

Table 1. Comparison Between the Amyloid Forming Regions Experimentally Discovered in Disease-Linked Polypeptides and the Stretches Predicted by Different Programs

Protein	Experimental regions	Predicted regions						
		SALSA [144]	Yoon and Welsh [147]	Galzitskaya *et al.* [149]	Nelson *et al.* [150]	Saiki *et al.* [153]	PASTA [154]	AGGRESCAN [160]

Aβ protein	17-21
	31-36
	38-42
		8-40*	11-13;
			16-20
				15-22	8-15;
					18-37
						13; 15;
						**17-19; 21;
						27; 31; 33; 35;
						37-38**
							**12-20;
							31-40**
								**17-22;
								30-42**

α-synuclein	31-109	**32-89**	no data available	no predicted regions	no data available	4-5; 13; 15-16;
						26;
						**38-39; 46; 48;
						50-53; 55; 61;
						64; 66; 68; 70;
						72-75; 77; 79;
						81; 85; 91; 95;**
						121
							**48-55;
							70-77**
								**36-42;
								49-55;
								87-94**

IAPP	8-20	no data available	**6-9; 14-20;** 25-30	**12-18**	no data available	**14; 16-17**	**12-32**	**13-18**

tau	301-320	120-130;
		220-230;
		296-326;
		390-410;
		412-430
			no data available	no predicted regions	11-17;
					27-31;
					61-64;
					100-106;
					116-150;
					162-167;
					254-267;
					273-287;
					294-299
					**305-310;
					316-319;**
					326-331;
					337-346;
					353-362;
					369-377;
					385-396;
					411-422;
					429-441
						no data available	**306-310**	**304-311**

*The predicted regions that match with the experimental results are indicated in bold.

Table 2. WWW Available Algorithms to Predict Relative Aggregation Rates and/or Amyloidogenic Segments of Protein Sequences.

Name	Availability	Reference
AGGRESCAN	http://bioinf.uab.es/aggrescan/	[160]
PAGE	caflisch@bioc.unizh.ch	[142]
PASTA	http://protein.cribi.unipd.it/pasta	[154]
SALSA	l.c.serpell@sussex.ac.uk	[144]
TANGO	http://tango.embl.de/	[143]
ZYGGREGATOR	http://www-vendruscolo.ch.cam.ac.uk/zyggregator.php	[141]

provide overlapping predictions on the determinants of aggregation in disease-linked polypeptides. In (Table **2**) the reader can find a list of World Wide Web available algorithms to predict relative aggregation rates and/or amyloidogenic segments of protein sequences.

In Silico Simulation of Protein Aggregation

In conformational diseases the appearance of pathological forms of proteins involves a complex aggregation pathway in which the protein can coexist in a number of different monomeric or oligomeric conformational ensembles, each with characteristic kinetic, thermodynamic and binding properties. Thus, the pathway can be viewed as a ragged free-energy landscape whose shape is delineated by the collection of protein conformations. Although the complete characterization of the different conformational states of a protein is a highly desirable goal and would provide the most accurate description of the pathway, the experimental means to achieve this objective are not obvious. Theoretically, one can obtain a picture of one of these particular states by applying techniques in the solid state (i.e., X-ray diffraction or SSNMR) or by NMR in solution. However, folding intermediates and oligomers are generally difficult to isolate due to their short half-life and highly flexible structures, and thus they are usually not amenable to these experimental approaches. Computational approaches and specifically molecular dynamics (MD) simulations might aid to fill this gap by providing insights, mechanistic explanations, and guiding experimental studies to tackle the problem of conformational diversity in pathological proteins. Importantly, the effects of specific changes in the polypeptide sequence or in the simulated environment can be also evaluated *in silico*. In addition, computational simulations might be of help in targeted drug design, e.g. by identifying chemical compounds aimed to interfere selective steps in the aggregation pathway. The areas of methodology development, simulations of aggregation mechanisms, and computational evaluation of amyloid structures have been the subject of excellent reviews [161,162]. Here, we will shortly review recent results on MD simulations of pathological protein systems and their possible use towards the development of therapeutic agents with the potential to treat conformational diseases.

All-atom MD simulations with explicit solvent are the most precise *in silico* approximations to model protein folding and assembly processes. These approaches employ high-resolution protein models based on a realistic representation of protein geometry and monitor continuously the positions and forces among all protein atoms along with their surrounding water molecules. They have been used mainly to explore conformational fluctuations in native or intermediate species and to analyze the stability and dynamics of particular proposals for the oligomeric or aggregated amyloid β-sheet structure. The potential of this approach is illustrated by a recent study by Thirumalai and co-workers in which extensive all-atom MD simulations were performed to understand how a preformed highly dynamic oligomeric assembly interacts with a nascent monomer. This is of great interest because metastable oligomers are being increasingly identified as the cytotoxic forms related to onset of disease. The authors studied the incorporation of a monomer of the CHC region of Aβ to well-defined preformed oligomers of different sizes formed by the same Aβ region [163]. The authors observed that when disordered monomer was added to an ordered oligomer, growth occurs largely by a two-phase dock-lock mechanism. The maximum change in the conformation of the monomer occurs during the rapid dock phase whereas in the much slower lock phase, the monomer forms a β-strand that is in register with the rest of the oligomer. An *a priori* unexpected observation was that, unlike during fibril growth, the initially ordered oligomer also partially disorders before forming a stable ordered structure. This dock and lock behavior could be a generic path for formation of toxic oligomers of amyloidogenic peptides and thus a target for therapeutic intervention.

Pure all-atom simulations are extremely demanding in computational terms because the length of a simulation is approximately proportional to the number of considered atoms to the third power. This complexity can be reduced by eliminating the solvent through its implicit consideration within the potential function governing the interactions between the different species in the system. This strategy was employed by Cheon and co-workers [164] to investigate the early stages of the oligomerization process for two fragments of the Aβ peptide: $A\beta_{16-22}$ (KLVFFAE) and $A\beta_{25-35}$ (GSNKGAIIGLM) respectively. The MD simulations were carried out with PROFASI [165], a Monte Carlo algorithm developed by Irback and co-workers, which implements an implicit water all-atom model, and included systems of 20 peptides for each Aβ fragment. The conversion of the monomeric peptides into oligomers could be rationalized into a process occurring through a generic two-step mechanism modulated by the competition between hydrogen bonding and peptide hydrophobicity (Fig. **5**). Depending on the balance of forces, the first step may be the coalescing of the peptides into more or less ordered oligomers. This step could be fast or even nonexistent, and, since hydrophobic forces drive it, it is nonspecific. The slower, second step comprises the reorganization of the peptides to form ordered oligomers, a process dependent on the formation of specific interchain hydrogen bonds by atoms in the peptide backbone, which results in the generation of β-sheets within the oligomeric species.

The kinetics of spontaneous amyloid fibril formation might take from seconds to days. The access to those time scales in the simulations can be only gained by sacrificing some detail in the models. In these simplified systems, beads with associated specific physicochemical properties represent the amino acids on the poylypeptide chain and, usually, accelerated methods are used to sample the resulting vast conformational space. This way, Caflisch and co-workers have reduced computational demand by using a coarse-grained model polypeptide composed of ten spherical beads disposed to have an overall amphipathic character and simulating 125 monomers in a cubic box [166]. They provide the monomer with internal flexibility and consider simplified free energy profiles with only two minima at the β-amyloid-competent state and the amyloid-protected state. The latter corresponds to the ensemble of conformers not compatible with the cross-β structure in a fibril. This strategy allows kinetic and thermodynamic analysis of the system. Interestingly, the results provide insights into the experimentally observed diversity of fibril formation mechanisms, suggesting that they depend on the stability of the amyloid-competent state of the monomer. It appears that in spite of the common final structure of the fibrils, high and low β-prone sequences encode for totally different aggregation kinetics. According to the model, fibrillogenesis of β-stable polypeptides would follow a downhill pathway without significant accumulation

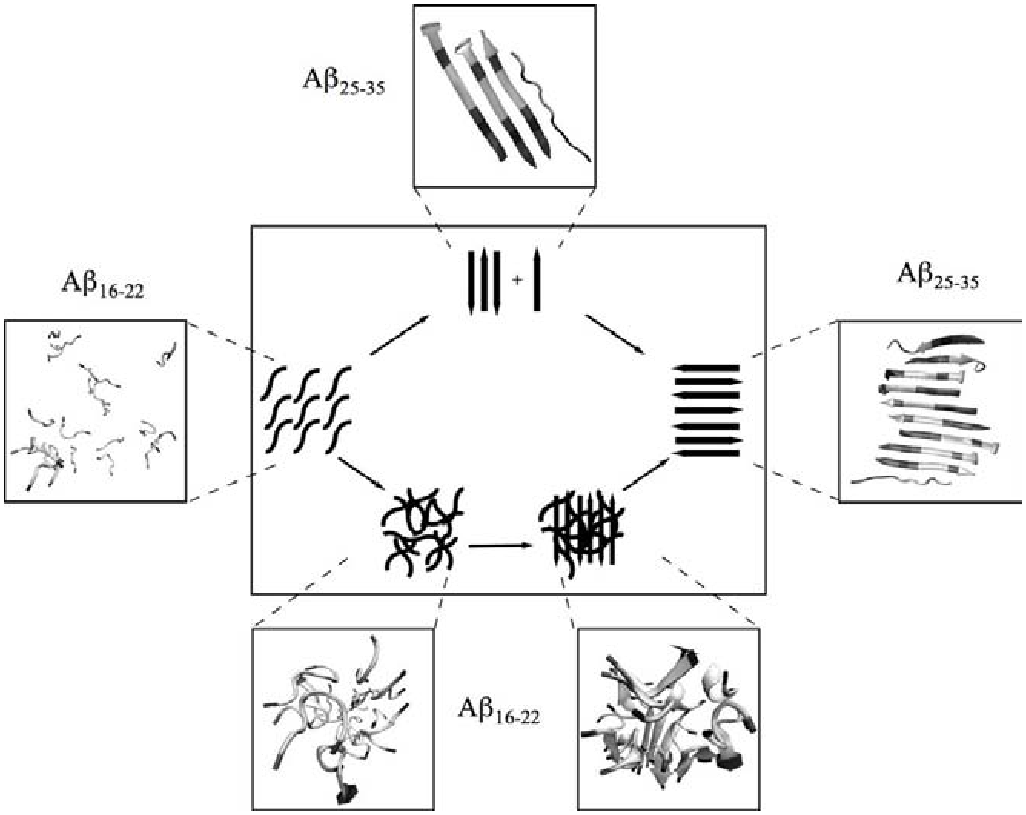


Fig. (5). General scheme of amyloid fibril formation pathways simulated by MD. The figure ilustrates the possible pathways that may follow peptide monomers to form amyloid fibrils. Representative structures of the simulated $A\beta_{25\text{-}35}$ (GSNKGAIIGLM) and $A\beta_{16\text{-}22}$ (KLVFFAE) fragments are shown. For $A\beta_{25\text{-}35}$ the formation of an ordered oligomer proceeds directly via selective backbone interchain hydrogen bonding. In contrast, for $A\beta_{16\text{-}22}$ the first assembly step is nonspecific, sustained by hydrophobic interactions, and leads to the formation of weakly ordered oligomers. A second slower step results in the reorganization of the oligomers through backbone hydrogen bonding to gain the β-sheet structure. (Adapted from Ref. [164]).

of intermediates, whereas aggregation of β-unstable sequences requires travelling through a series of on-pathway oligomeric states in the lag phase, before fibril elongation occurs. It follows that, in this latter case, the presence of metastable off-pathway assemblies can compete with and decrease the likelihood of on-pathway interactions that would finally promote fibril formation. Still, even in this situation, fibrillar structures would represent a lower free energy state than metastable off-pathway oligomers, thus accounting for the thermodynamic bias towards the population of amyloidogenic pathways. Overall, these results imply that different strategies should be used to block the growth of fibrils depending on the nature of the polypeptide.

In general, despite their simplicity, coarse-grained approaches using minimalist models have allowed the construction of free-energy landscapes for protein aggregation processes. Nevertheless, although they are well suited to delineate general rules, they can hardly be used to address specific problems where selective backbone and side chains interactions modulate fibrillogenesis. This task can be undertaken with intermediate-resolution models containing explicit representations of the protein backbone and side chains. This way, Hall and co-workers developed the PRIME model [167,168], which represents each amino acid with four beads, three for the backbone and one for the side chain. This allows the treatment of large multichain systems while maintaining a fairly realistic description of protein dynamics. They combined the model with discontinuous molecular dynamics, a fast sampling procedure that is applicable to systems of molecules interacting via discontinuous potentials. Solvent was modeled implicitly and backbone hydrogen bonding in explicit detail. Overall, the approach allowed sampling much wider regions of conformational space, longer time scales, and larger systems than in traditional molecular dynamics. They used polyalanine peptides to benchmark the approach and found that small amorphous aggregates populated the pathway before the formation of the critical nucleus [169]. After nucleation, fibril growth depends on β-sheet elongation by the addition of monomers to the end of each β-sheet and on lateral extension by incorporation of preformed β-sheets.

One of the most promising applications of MD to the protein aggregation field is the possibility to design chemical compounds against specific intermediate conformers in the fibrillogenesis pathway. In the case of AD, the structural complexity of Aβ is a major problem for the design of chemical compounds that bind to it specifically. Importantly, for Aβ, the group of Nussinov has shown that computational simulations might become a powerful discovery tool, since their approach [170] was able to foresee with great detail the solid-NMR structure of the fibrillar state of the peptide [171]. The formation of Aβ fibrils *in vivo* is thought to be mediated by the conformational transition of Aβ from α-helix to β-sheet or from random coil to β-sheet. According to various experimental assays, Aβ may adopt multiple conformations *in vitro*, such as α-helices, β-sheets, or random coils. These conformations depend on the buffer conditions of pH, ionic strength, and solvent properties. Although some chemical compounds have been demonstrated to reduce the cytotoxicity of Aβ peptides, in most cases it is not clear to which Aβ conformation or assembly they bind. The development of toxic conformation-directed drugs to treat AD is especially attractive because of their predicted specificity and low toxicity. MD simulations on the full $A\beta_{1\text{-}40}$ peptide by Jiang and co-workers showed that the conformational transition from α-helix to random coil passes through an α-helix/β-sheet intermediate structure. Furthermore, it was observed that the α-helix/β-sheet intermediate structure possesses a core domain, within which four glycine residues (G25, G29, G33, and G37) are essential to β-sheet formation and amyloid fibrillogenesis. The authors proposed that chemical compounds that lock the structure of Aβ into the α-helix/β-sheet intermediate state could potentially inhibit Aβ fibrillogenesis [172]. Under this hypothesis, virtual screening based on molecular docking was performed, targeting an Aβ peptide α-helix/β-sheet intermediate structure extracted from the trajectory of a 50-ns MD simulation on $A\beta_{1-40}$. A commercial database of small-molecule compounds was queried using DOCK4.0 (http://dock.compbio.ucsf.edu/) as the search engine. The top 1,000 ranked entities according to the energy score were re-evaluated, and the interaction with $A\beta_{1-40}$ was carefully inspected. More than hundred candidate compounds were used in biological assays. One of them inhibited $A\beta_{1-42}$ fibrillogenesis in a concentration-dependent manner. Results from several assays led to the proposal that the inhibitor stabilized the β-sheet conformation of Aβ. Modelling suggested that the inhibitor binds to the target mainly through hydrophobic interactions [173]. This work offers a proof of principle that computational simulations might contribute to render valuable chemicals against conformational diseases.

Future Directions

The information contained in the present review clearly demonstrates that both structural and computational approaches to the investigation of protein aggregation associated with conformational diseases are undergoing a rapid and very productive phase of growth. In the past, both kinds of analysis have been performed largely separated from each other and only recently has information exchange between wet and dry experimental approaches begun to occur. It is very likely that in the next few years, we will witness the construction of a highly synergic environment in which both kinds of data are integrated to attain an unambiguous and accurate description of the mechanism underlying protein aggregation, the conformers eliciting the cytotoxic effect and the way these processes are regulated within the cell. The knowledge gained at this basic research level might well be translated into novel and effective alternative therapeutics addressed to the treatment of Alzheimer's, Parkinson's and other debilitating diseases caused by protein misfolding and deposition.

ACKNOWLEDGEMENTS

Space limitation precludes citation of a large body of excellent work by colleagues in the fields discussed. We acknowledge these efforts here. This work was supported by Grants CSD2006-00012 from the Ministerio de Educación y Ciencia, Spain, which included Fondo Europeo de Desarrollo Regional funds, BIO2007-68046 and SGR00037 (Generalitat de Catalunya).

ABBREVIATIONS

3D	=	Three-dimensional
Aβ	=	Amyloid-β protein
AD	=	Alzheimer's disease
AFM	=	Atomic force microscopy
β2m	=	β2 microglobulin
CHC	=	Central hydrophobic cluster
Cryo-TEM	=	Cryogenic transmission electron microscopy
FCS	=	Fluorescence correlation spectroscopy
GFP	=	Green fluorescent protein
IAPP	=	Islet amyloid polypeptide
Ig	=	Immunoglobulin
MD	=	Molecular dynamics
NMR	=	Nuclear magnetic resonance
NSOM	=	Near-field scanning optical microscopy
PASTA	=	Prediction of amyloid structure aggregation
PDB	=	Brookhaven protein data bank
polyQ	=	Polyglutamine
RNAse A	=	Ribonuclease A
SALSA	=	Single algorithm for sliding averages
SH3	=	Src homology 3
SMFS	=	Single molecule force spectroscopy
SPM	=	Scanning probe microscopy
SSNMR	=	Solid-state nuclear magnetic resonance
STEM	=	Scanning transmission electron microscopy
STM	=	Scanning tunneling microscopy
TEM	=	Transmission electron microscopy
ThT	=	Thioflavin T
TIRFM	=	Total internal reflection fluorescence microscopy
TTR	=	Transthyretin

REFERENCES

[1] Luheshi, L. M.; Tartaglia, G. G.; Brorsson, A. C.; Pawar, A. P.; Watson, I. E.; Chiti, F.; Vendruscolo, M.; Lomas, D. A.; Dobson, C. M.; Crowther, D. C. *PLoS Biol.,* **2007**, *5*, e290.

[2] Chiti, F.; Dobson, C. M. *Annu. Rev. Biochem.,* **2006**, *75*, 333.

[3] Dobson, C. M. *Trends Biochem. Sci.,* **1999**, *24*, 329.

[4] Fandrich, M.; Dobson, C. M. *EMBO J.,* **2002**, *21*, 5682.

[5] Huff, M. E.; Balch, W. E.; Kelly, J. W. *Curr. Opin. Struct. Biol.,* **2003**, *13*, 674.

[6] Lomas, D. A.; Carrell, R. W. *Nat. Rev. Genet.,* **2002**, *3*, 759.

[7] Selkoe, D. J. *Nature,* **2003**, *426*, 900.

[8] Bossy-Wetzel, E.; Schwarzenbacher, R.; Lipton, S. A. *Nat. Med.,* **2004**, *10 Suppl*, S2.

[9] Stefani, M. *Biochim. Biophys. Acta,* **2004**, *1739*, 5.

[10] Chiti, F.; Dobson, C. M. *Annu. Rev. Biochem.,* **2006**, *75*, 333.

[11] Petkova, A. T.; Ishii, Y.; Balbach, J. J.; Antzutkin, O. N.; Leapman, R. D.; Delaglio, F.; Tycko, R. *Proc. Natl. Acad. Sci. U.S.A,* **2002**, *99*, 16742.

[12] Jaroniec, C. P.; MacPhee, C. E.; Astrof, N. S.; Dobson, C. M.; Griffin, R. G. *Proc. Natl. Acad. Sci. U.S.A,* **2002**, *99*, 16748.

[13] Ritter, C.; Maddelein, M. L.; Siemer, A. B.; Luhrs, T.; Ernst, M.; Meier, B. H.; Saupe, S. J.; Riek, R. *Nature,* **2005**, *435*, 844.

[14] Makin, O. S.; Serpell, L. C. *FEBS J.,* **2005**, *272*, 5950.

[15] Nelson, R.; Sawaya, M. R.; Balbirnie, M.; Madsen, A. O.; Riekel, C.; Grothe, R.; Eisenberg, D. *Nature,* **2005**, *435*, 773.

[16] Makin, O. S.; Atkins, E.; Sikorski, P.; Johansson, J.; Serpell, L. C. *Proc. Natl. Acad. Sci. U.S.A,* **2005**, *102*, 315.

[17] Makin, O. S.; Serpell, L. C. *Biochem. Soc. Trans.,* **2002**, *30*, 521.

[18] Uversky, V. N.; Kabanov, A. V.; Lyubchenko, Y. L. *J. Proteome Res.,* **2006**, *5*, 2505.

[19] Baldus, M. *Angew. Chem. Int. Ed. Engl.,* **2006**, *45*, 1186.

[20] Tycko, R. *Methods Enzymol.,* **2006**, *413*, 103.

[21] Tycko, R. *Protein Pept. Lett.,* **2006**, *13*, 229.

[22] Naito, A.; Kawamura, I. *Biochim. Biophys. Acta,* **2007**, *1768*, 1900.

[23] Lansbury, P. T., Jr.; Costa, P. R.; Griffiths, J. M.; Simon, E. J.; Auger, M.; Halverson, K. J.; Kocisko, D. A.; Hendsch, Z. S.; Ashburn, T. T.; Spencer, R. G. *Nat. Struct. Biol.,* **1995**, *2*, 990.

[24] Benzinger, T. L.; Gregory, D. M.; Burkoth, T. S.; Miller-Auer, H.; Lynn, D. G.; Botto, R. E.; Meredith, S. C. *Biochemistry,* **2000**, *39*, 3491.

[25] Antzutkin, O. N.; Leapman, R. D.; Balbach, J. J.; Tycko, R. *Biochemistry,* **2002**, *41*, 15436.

[26] Kajava, A. V.; Baxa, U.; Wickner, R. B.; Steven, A. C. *Proc. Natl. Acad. Sci. U.S.A,* **2004**, *101*, 7885.

[27] Wille, H.; Michelitsch, M. D.; Guenebaut, V.; Supattapone, S.; Serban, A.; Cohen, F. E.; Agard, D. A.; Prusiner, S. B. *Proc. Natl. Acad. Sci. U.S.A,* **2002**, *99*, 3563.

[28] Guo, J. T.; Wetzel, R.; Xu, Y. *Proteins,* **2004**, *57*, 357.

[29] Diaz-Avalos, R.; Long, C.; Fontano, E.; Balbirnie, M.; Grothe, R.; Eisenberg, D.; Caspar, D. L. *J. Mol. Biol.,* **2003**, *330*, 1165.

[30] Antzutkin, O. N.; Balbach, J. J.; Leapman, R. D.; Rizzo, N. W.; Reed, J.; Tycko, R. *Proc. Natl. Acad. Sci. U.S.A,* **2000**, *97*, 13045.

[31] Balbach, J. J.; Petkova, A. T.; Oyler, N. A.; Antzutkin, O. N.; Gordon, D. J.; Meredith, S. C.; Tycko, R. *Biophys. J.,* **2002**, *83*, 1205.

[32] Antzutkin, O. N.; Balbach, J. J.; Tycko, R. *Biophys. J.,* **2003**, *84*, 3326.

[33] Petkova, A. T.; Leapman, R. D.; Guo, Z.; Yau, W. M.; Mattson, M. P.; Tycko, R. *Science,* **2005**, *307*, 262.

[34] Oyler, N. A.; Tycko, R. *J. Am. Chem. Soc.,* **2004**, *126*, 4478.

[35] Balbach, J. J.; Ishii, Y.; Antzutkin, O. N.; Leapman, R. D.; Rizzo, N. W.; Dyda, F.; Reed, J.; Tycko, R. *Biochemistry,* **2000**, *39*, 13748.

[36] Tycko, R.; Ishii, Y. *J. Am. Chem. Soc.,* **2003**, *125*, 6606.

[37] Petkova, A. T.; Buntkowsky, G.; Dyda, F.; Leapman, R. D.; Yau, W. M.; Tycko, R. *J. Mol. Biol.,* **2004**, *335*, 247.

[38] Luhrs, T.; Ritter, C.; Adrian, M.; Riek-Loher, D.; Bohrmann, B.; Dobeli, H.; Schubert, D.; Riek, R. *Proc. Natl. Acad. Sci. U.S.A,* **2005**, *102*, 17342.

[39] Jaroniec, C. P.; MacPhee, C. E.; Bajaj, V. S.; McMahon, M. T.; Dobson, C. M.; Griffin, R. G. *Proc. Natl. Acad. Sci. U.S.A,* **2004**, *101*, 711.

[40] Kammerer, R. A.; Kostrewa, D.; Zurdo, J.; Detken, A.; Garcia-Echeverria, C.; Green, J. D.; Muller, S. A.; Meier, B. H.; Winkler, F. K.; Dobson, C. M.; Steinmetz, M. O. *Proc. Natl. Acad. Sci. U.S.A,* **2004**, *101*, 4435.

[41] Laws, D. D.; Bitter, H. M.; Liu, K.; Ball, H. L.; Kaneko, K.; Wille, H.; Cohen, F. E.; Prusiner, S. B.; Pines, A.; Wemmer, D. E. *Proc. Natl. Acad. Sci. U.S.A,* **2001**, *98*, 11686.

[42] Kawahara, M.; Kuroda, Y. *Brain Res. Bull.,* **2000**, *53*, 389.

[43] Kayed, R.; Sokolov, Y.; Edmonds, B.; McIntire, T. M.; Milton, S. C.; Hall, J. E.; Glabe, C. G. *J. Biol. Chem.,* **2004**, *279*, 46363.

[44] Lin, M. C.; Kagan, B. L. *Peptides,* **2002**, *23*, 1215.

[45] Ravault, S.; Soubias, O.; Saurel, O.; Thomas, A.; Brasseur, R.; Milon, A. *Protein Sci.,* **2005**, *14*, 1181.

[46] Neto-Silva, R. M.; Macedo-Ribeiro, S.; Pereira, P. J.; Coll, M.; Saraiva, M. J.; Damas, A. M. *Acta Crystallogr. D Biol. Crystallogr.,* **2005**, *61*, 333.

[47] Rosano, C.; Zuccotti, S.; Bolognesi, M. *Biochim. Biophys. Acta,* **2005**, *1753*, 85.

[48] Goldsbury, C. S.; Wirtz, S.; Muller, S. A.; Sunderji, S.; Wicki, P.; Aebi, U.; Frey, P. *J. Struct. Biol.,* **2000**, *130*, 217.

[49] Chamberlain, A. K.; MacPhee, C. E.; Zurdo, J.; Morozova-Roche, L. A.; Hill, H. A.; Dobson, C. M.; Davis, J. J. *Biophys. J.,* **2000**, *79*, 3282.

[50] Arimon, M.; Diez-Perez, I.; Kogan, M. J.; Durany, N.; Giralt, E.; Sanz, F.; Fernandez-Busquets, X. *FASEB J.,* **2005**, *19*, 1344.

[51] Ruben, G. C.; Wang, J. Z.; Iqbal, K.; Grundke-Iqbal, I. *Microsc. Res. Tech.,* **2005**, *67*, 175.

[52] Serpell, L. C.; Smith, J. M. *J. Mol. Biol.,* **2000**, *299*, 225.

[53] Jimenez, J. L.; Guijarro, J. I.; Orlova, E.; Zurdo, J.; Dobson, C. M.; Sunde, M.; Saibil, H. R. *EMBO J.,* **1999**, *18*, 815.

[54] Jimenez, J. L.; Tennent, G.; Pepys, M.; Saibil, H. R. *J. Mol. Biol.,* **2001**, *311*, 241.

[55] Sumner, M. O.; Serpell, L. C. *J. Mol. Biol.,* **2004**, *335*, 1279.

[56] Tycko, R. *Curr. Opin. Struct. Biol.,* **2004**, *14*, 96.

[57] Kajava, A. V.; Aebi, U.; Steven, A. C. *J. Mol. Biol.,* **2005**, *348*, 247.

[58] Goldsbury, C. S.; Cooper, G. J.; Goldie, K. N.; Muller, S. A.; Saafi, E. L.; Gruijters, W. T.; Misur, M. P.; Engel, A.; Aebi, U.; Kistler, J. *J. Struct. Biol.,* **1997**, *119*, 17.

[59] Serpell, L. C.; Sunde, M.; Fraser, P. E.; Luther, P. K.; Morris, E. P.; Sangren, O.; Lundgren, E.; Blake, C. C. *J. Mol. Biol.,* **1995**, *254*, 113.

[60] Tattum, M. H.; Cohen-Krausz, S.; Thumanu, K.; Wharton, C. W.; Khalili-Shirazi, A.; Jackson, G. S.; Orlova, E. V.; Collinge, J.; Clarke, A. R.; Saibil, H. R. *J. Mol. Biol.,* **2006**, *357*, 975.

[61] Jimenez, J. L.; Nettleton, E. J.; Bouchard, M.; Robinson, C. V.; Dobson, C. M.; Saibil, H. R. *Proc. Natl. Acad. Sci. U.S.A,* **2002**, *99*, 9196.

[62] Dobson, C. M. *Nature,* **2003**, *426*, 884.

[63] Collins, S. R.; Douglass, A.; Vale, R. D.; Weissman, J. S. *PLoS Biol.,* **2004**, *2*, e321.

[64] Tanaka, M.; Collins, S. R.; Toyama, B. H.; Weissman, J. S. *Nature,* **2006**, *442*, 585.

[65] Cornish, P. V.; Ha, T. *ACS Chem. Biol.,* **2007**, *2*, 53.

[66] Lyubchenko, Y. L.; Sherman, S.; Shlyakhtenko, L. S.; Uversky, V. N. *J. Cell Biochem.,* **2006**, *99*, 52.

[67] Binnig, G.; Quate, C. F.; Gerber, C. *Phys. Rev. Lett.,* **1986**, *56*, 930.

[68] Poggi, M. A.; Gadsby, E. D.; Bottomley, L. A.; King, W. P.; Oroudjev, E.; Hansma, H. *Anal. Chem.,* **2004**, *76*, 3429.

[69] Engel, A.; Muller, D. J. *Nat. Struct. Biol.,* **2000**, *7*, 715.

[70] Horber, J. K.; Miles, M. J. *Science,* **2003**, *302*, 1002.

[71] Shivji, A. P.; Brown, F.; Davies, M. C.; Jennings, K. H.; Roberts, C. J.; Tendler, S. J.; Wilkinson, M. J.; Williams, P. M. *FEBS Lett.,* **1995**, *371*, 25.

[72] Wang, Z.; Zhou, C.; Wang, C.; Wan, L.; Fang, X.; Bai, C. *Ultramicroscopy,* **2003**, *97*, 73.

[73] Losic, D.; Martin, L. L.; Mechler, A.; Aguilar, M. I.; Small, D. H. *J. Struct. Biol.,* **2006**, *155*, 104.

[74] Fritz, J.; Anselmetti, D.; Jarchow, J.; Fernandez-Busquets, X. *J. Struct. Biol.,* **1997**, *119*, 165.

[75] Gosal, W. S.; Myers, S. L.; Radford, S. E.; Thomson, N. H. *Protein Pept. Lett.,* **2006**, *13*, 261.

[76] Kowalewski, T.; Holtzman, D. M. *Proc. Natl. Acad. Sci. U.S.A,* **1999**, *96*, 3688.

[77] Khurana, R.; Ionescu-Zanetti, C.; Pope, M.; Li, J.; Nielson, L.; Ramirez-Alvarado, M.; Regan, L.; Fink, A. L.; Carter, S. A. *Biophys. J.,* **2003**, *85*, 1135.

[78] Hoyer, W.; Cherny, D.; Subramaniam, V.; Jovin, T. M. *J. Mol. Biol.,* **2004**, *340*, 127.

[79] Kad, N. M.; Myers, S. L.; Smith, D. P.; Smith, D. A.; Radford, S. E.; Thomson, N. H. *J. Mol. Biol.,* **2003**, *330*, 785.

[80] Goldsbury, C.; Kistler, J.; Aebi, U.; Arvinte, T.; Cooper, G. J. *J. Mol. Biol.,* **1999**, *285*, 33.

[81] Stolz, M.; Stoffler, D.; Aebi, U.; Goldsbury, C. *J. Struct. Biol.,* **2000**, *131*, 171.

[82] Zlatanova, J.; Lindsay, S. M.; Leuba, S. H. *Prog. Biophys. Mol. Biol.,* **2000**, *74*, 37.

[83] Evans, E.; Ritchie, K. *Biophys. J.,* **1997**, *72*, 1541.

[84] Merkel, R.; Nassoy, P.; Leung, A.; Ritchie, K.; Evans, E. *Nature,* **1999**, *397*, 50.

[85] Hoh, J. H.; Cleveland, J. P.; Prater, C. B.; Revel, J. P.; Hansma, P. K. *J. Am. Chem. Soc.,* **1992**, *114*, 4917.

[86] Kellermayer, M. S.; Grama, L.; Karsai, A.; Nagy, A.; Kahn, A.; Datki, Z. L.; Penke, B. *J. Biol. Chem.,* **2005**, *280*, 8464.

[87] Kransnoslobodtsev, A. V.; Shlyakhtenko, L. S.; Ukraintsev, E.; Zaikova, T. O.; Keana, J. F.; Lyubchenko, Y. L. *Nanomedicine,* **2005**, *1*, 300.

[88] McAllister, C.; Karymov, M. A.; Kawano, Y.; Lushnikov, A. Y.; Mikheikin, A.; Uversky, V. N.; Lyubchenko, Y. L. *J. Mol. Biol.,* **2005**, *354*, 1028.

[89] Smith, J. F.; Knowles, T. P.; Dobson, C. M.; MacPhee, C. E.; Welland, M. E. *Proc. Natl. Acad. Sci. U.S.A,* **2006**, *103*, 15806.

[90] Kessler, M.; Gaub, H. E. *Structure,* **2006**, *14*, 521.

[91] Garcia-Manyes, S.; Bucior, I.; Ros, R.; Anselmetti, D.; Sanz, F.; Burger, M. M.; Fernandez-Busquets, X. *J. Biol. Chem.,* **2006**, *281*, 5992.

[92] Pepys, M. B. *Philos. Trans. R. Soc. Lond. B Biol. Sci.,* **2001**, *356*, 203.

[93] Leapman, R. D.; Rizzo, N. W. *Ultramicroscopy,* **1999**, *78*, 251.

[94] Ban, T.; Goto, Y. *Methods Enzymol.,* **2006**, *413*, 91.

[95] Wazawa, T.; Ueda, M. *Adv. Biochem. Eng. Biotechnol.,* **2005**, *95*, 77.

[96] Inoue, Y.; Kishimoto, A.; Hirao, J.; Yoshida, M.; Taguchi, H. *J. Biol. Chem.,* **2001**, *276*, 35227.

[97] Ban, T.; Hamada, D.; Hasegawa, K.; Naiki, H.; Goto, Y. *J. Biol. Chem.,* **2003**, *278*, 16462.

[98] Alexandrescu, A. T. *Protein Sci.,* **2005**, *14*, 1.

[99] Widenbrant, M. J.; Rajadas, J.; Sutardja, C.; Fuller, G. G. *Biophys. J.,* **2006**, *91*, 4071.

[100] Murphy, R. M. *Biochim. Biophys. Acta,* **2007**, *1768*, 1923.

[101] Munishkina, L. A.; Fink, A. L. *Biochim. Biophys. Acta,* **2007**, *1768*, 1862.

[102] de Lange, F.; Cambi, A.; Huijbens, R.; de Bakker, B.; Rensen, W.; Garcia-Parajo, M.; van Hulst, N.; Figdor, C. G. *J. Cell Sci.,* **2001**, *114*, 4153.

[103] Levin, M. K.; Carson, J. H. *Differentiation,* **2004**, *72*, 1.

[104] Hess, S. T.; Huang, S.; Heikal, A. A.; Webb, W. W. *Biochemistry,* **2002**, *41*, 697.

[105] Tjernberg, L. O.; Pramanik, A.; Bjorling, S.; Thyberg, P.; Thyberg, J.; Nordstedt, C.; Berndt, K. D.; Terenius, L.; Rigler, R. *Chem. Biol.,* **1999**, *6*, 53.

[106] Hossain, S.; Grande, M.; Ahmadkhanov, G.; Pramanik, A. *Exp. Mol. Pathol.,* **2007**, *82*, 169.

[107] Verdier, Y.; Zarandi, M.; Penke, B. *J. Pept. Sci.,* **2004**, *10*, 229.

[108] Boland, K.; Behrens, M.; Choi, D.; Manias, K.; Perlmutter, D. H. *J. Biol. Chem.,* **1996**, *271*, 18032.

[109] Matter, M. L.; Zhang, Z.; Nordstedt, C.; Ruoslahti, E. *J. Cell Biol.,* **1998**, *141*, 1019.

[110] Bi, X.; Gall, C. M.; Zhou, J.; Lynch, G. *Neuroscience,* **2002**, *112*, 827-840.

[111] Wang, H. Y.; Li, W.; Benedetti, N. J.; Lee, D. H. *J. Biol. Chem.,* **2003**, *278*, 31547.

[112] Du Yan, S.; Zhu, H.; Fu, J.; Yan, S. F.; Roher, A.; Tourtellotte, W. W.; Rajavashisth, T.; Chen, X.; Godman, G. C.; Stern, D.; Schmidt, A. M. *Proc. Natl. Acad. Sci. U.S.A,* **1997**, *94*, 5296.

[113] Husemann, J.; Loike, J. D.; Anankov, R.; Febbraio, M.; Silverstein, S. C. *Glia,* **2002**, *40*, 195.

[114] Coraci, I. S.; Husemann, J.; Berman, J. W.; Hulette, C.; Dufour, J. H.; Campanella, G. K.; Luster, A. D.; Silverstein, S. C.; El-Khoury, J. B. *Am. J. Pathol.,* **2002**, *160*, 101.

[115] Le, Y.; Gong, W.; Tiffany, H. L.; Tumanov, A.; Nedospasov, S.; Shen, W.; Dunlop, N. M.; Gao, J. L.; Murphy, P. M.; Oppenheim, J. J.; Wang, J. M. *J. Neurosci.,* **2001**, *21*, RC123.

[116] Dineley, K. T.; Bell, K. A.; Bui, D.; Sweatt, J. D. *J. Biol. Chem.,* **2002**, *277*, 25056.

[117] Xie, L.; Helmerhorst, E.; Taddei, K.; Plewright, B.; Van Bronswijk, W.; Martins, R. *J. Neurosci.,* **2002**, *22*, RC221.

[118] Yaar, M.; Zhai, S.; Fine, R. E.; Eisenhauer, P. B.; Arble, B. L.; Stewart, K. B.; Gilchrest, B. A. *J. Biol. Chem.,* **2002**, *277*, 7720.

[119] Pitschke, M.; Prior, R.; Haupt, M.; Riesner, D. *Nat. Med.,* **1998**, *4*, 832.

[120] Takahashi, Y.; Okamoto, Y.; Popiel, H. A.; Fujikake, N.; Toda, T.; Kinjo, M.; Nagai, Y. *J. Biol. Chem.,* **2007**, *282*, 24039.

[121] Sunde, M.; Serpell, L. C.; Bartlam, M.; Fraser, P. E.; Pepys, M. B.; Blake, C. C. *J. Mol. Biol.,* **1997**, *273*, 729.

[122] Eisenberg, D.; Nelson, R.; Sawaya, M. R.; Balbirnie, M.; Sambashivan, S.; Ivanova, M. I.; Madsen, A. O.; Riekel, C. *Acc. Chem. Res.,* **2006**, *39*, 568.

[123] Bousset, L.; Briki, F.; Doucet, J.; Melki, R. *J. Struct. Biol.,* **2003**, *141*, 132.

[124] Baxa, U.; Taylor, K. L.; Wall, J. S.; Simon, M. N.; Cheng, N.; Wickner, R. B.; Steven, A. C. *J. Biol. Chem.,* **2003**, *278*, 43717.

[125] Inouye, H.; Domingues, F. S.; Damas, A. M.; Saraiva, M. J.; Lundgren, E.; Sandgren, O.; Kirschner, D. A. *Amyloid,* **1998**, *5*, 163.

[126] Eakin, C. M.; Attenello, F. J.; Morgan, C. J.; Miranker, A. D. *Biochemistry,* **2004**, *43*, 7808.

[127] Lazo, N. D.; Downing, D. T. *Biochem. Biophys. Res. Commun.,* **1997**, *235*, 675.

[128] Lazo, N. D.; Downing, D. T. *Biochemistry,* **1998**, *37*, 1731.

[129] Perutz, M. F.; Finch, J. T.; Berriman, J.; Lesk, A. *Proc. Natl. Acad. Sci. U.S.A,* **2002**, *99*, 5591.

[130] Wetzel, R. *Structure,* **2002**, *10*, 1031.

[131] Kheterpal, I.; Zhou, S.; Cook, K. D.; Wetzel, R. *Proc. Natl. Acad. Sci. U.S.A,* **2000**, *97*, 13597.

[132] Williams, A. D.; Portelius, E.; Kheterpal, I.; Guo, J. T.; Cook, K. D.; Xu, Y.; Wetzel, R. *J. Mol. Biol.,* **2004**, *335*, 833.

[133] Kishimoto, A.; Hasegawa, K.; Suzuki, H.; Taguchi, H.; Namba, K.; Yoshida, M. *Biochem. Biophys. Res. Commun.,* **2004**, *315*, 739.

[134] Malinchik, S. B.; Inouye, H.; Szumowski, K. E.; Kirschner, D. A. *Biophys. J.,* **1998**, *74*, 537.

[135] Kirschner, D. A.; Inouye, H.; Duffy, L. K.; Sinclair, A.; Lind, M.; Selkoe, D. J. *Proc. Natl. Acad. Sci. U.S.A,* **1987**, *84*, 6953.

[136] Fraser, P. E.; Duffy, L. K.; O'Malley, M. B.; Nguyen, J.; Inouye, H.; Kirschner, D. A. *J. Neurosci. Res.,* **1991**, *28*, 474.

[137] Inouye, H.; Fraser, P. E.; Kirschner, D. A. *Biophys. J.,* **1993**, *64*, 502.

[138] Carulla, N.; Caddy, G. L.; Hall, D. R.; Zurdo, J.; Gairi, M.; Feliz, M.; Giralt, E.; Robinson, C. V.; Dobson, C. M. *Nature,* **2005**, *436*, 554.

[139] Baskakov, I. V.; Bocharova, O. V. *Biochemistry,* **2005**, *44*, 2339.

[140] Serpell, L. C.; Sunde, M.; Benson, M. D.; Tennent, G. A.; Pepys, M. B.; Fraser, P. E. *J. Mol. Biol.,* **2000**, *300*, 1033.

[141] Pawar, A.P.; Dubay, K.F.; Zurdo, J.; Chiti, F.; Vendruscolo, M.; Dobson, C.M. *J. Mol. Biol.*, **2005**, *350,* 379.

[142] Caflisch, A. *Curr. Opin. Chem. Biol.,* **2006**, *10*, 437.

[143] Rousseau, F.; Schymkowitz, J.; Serrano, L. *Curr. Opin. Struct. Biol.,* **2006**, *16*, 118.

[144] Zibaee, S.; Makin, O. S.; Goedert, M.; Serpell, L. C. *Protein Sci.,* **2007**, *16*, 906.

[145] Chou, P. Y.; Fasman, G. D. *Adv. Enzymol. Relat. Areas Mol. Biol.,* **1978**, *47*, 45.

[146] Chiti, F.; Stefani, M.; Taddei, N.; Ramponi, G.; Dobson, C. M. *Nature,* **2003**, *424*, 805.

[147] Yoon, S.; Welsh, W. J. *Protein Sci.,* **2004**, *13*, 2149.

[148] Yoon, S.; Welsh, W. J. *Proteins,* **2005**, *60*, 110.

[149] Galzitskaya, O. V.; Garbuzynskiy, S. O.; Lobanov, M. Y. *PLoS Comput. Biol.,* **2006**, *2*, e177.

[150] Nelson, R.; Eisenberg, D. *Curr. Opin. Struct. Biol.,* **2006**, *16*, 260.

[151] Thompson, M. J.; Sievers, S. A.; Karanicolas, J.; Ivanova, M. I.; Baker, D.; Eisenberg, D. *Proc. Natl. Acad. Sci. U.S.A,* **2006**, *103*, 4074.

[152] Kuhlman, B.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A,* **2000**, *97*, 10383.

[153] Saiki, M.; Konakahara, T.; Morii, H. *Biochem. Biophys. Res. Commun.,* **2006**, *343*, 1262.

[154] Trovato, A.; Chiti, F.; Maritan, A.; Seno, F. *PLoS Comput. Biol.,* **2006**, *2*, e170.

[155] Rojas Quijano, F. A.; Morrow, D.; Wise, B. M.; Brancia, F. L.; Goux, W. J. *Biochemistry,* **2006**, *45*, 4638.

[156] Ma, Y.; Hendershot, L. M. *Cell,* **2001**, *107*, 827.

[157] Waldo, G. S.; Standish, B. M.; Berendzen, J.; Terwilliger, T. C. *Nat. Biotechnol.,* **1999**, *17*, 691.

[158] de Groot, N. S.; Aviles, F. X.; Vendrell, J.; Ventura, S. *FEBS J.,* **2006**, *273*, 658.

[159] de Groot, N.; Pallares, I.; Aviles, F.; Vendrell, J.; Ventura, S. *BMC Struct. Biol.,* **2005**, *5*, 18.

[160] Conchillo-Sole, O.; de Groot, N. S.; Aviles, F. X.; Vendrell, J.; Daura, X.; Ventura, S. *BMC Bioinformatics,* **2007**, *8*, 65.

[161] Cellmer, T.; Bratko, D.; Prausnitz, J. M.; Blanch, H. W. *Trends Biotechnol.,* **2007**, *25*, 254.

[162] Ma, B.; Nussinov, R. *Curr. Opin. Chem. Biol.,* **2006**, *10*, 445.

[163] Nguyen, P. H.; Li, M. S.; Stock, G.; Straub, J. E.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A,* **2007**, *104*, 111.

[164] Cheon, M.; Chang, I.; Mohanty, S.; Luheshi, L. M.; Dobson, C. M.; Vendruscolo, M.; Favrin, G. *PLoS Comput. Biol.,* **2007**, *3*, 1727.

[165] Irback, A.; Mohanty, S. *J. Comput. Chem.,* **2006**, *27*, 1548.

[166] Pellarin, R.; Caflisch, A. *J. Mol. Biol.,* **2006**, *360*, 882.

[167] Smith, A. V.; Hall, C. K. *J. Mol. Biol.,* **2001**, *312*, 187.

[168] Nguyen, H. D.; Marchut, A. J.; Hall, C. K. *Protein Sci.,* **2004**, *13*, 2909.

[169] Nguyen, H. D.; Hall, C. K. *Proc. Natl. Acad. Sci. U.S.A,* **2004**, *101*, 16180.

[170] Ma, B.; Nussinov, R. *Proc. Natl. Acad. Sci. U.S.A,* **2002**, *99*, 14126.

[171] Petkova, A. T.; Yau, W. M.; Tycko, R. *Biochemistry,* **2006**, *45*, 498.

[172] Xu, Y.; Shen, J.; Luo, X.; Zhu, W.; Chen, K.; Ma, J.; Jiang, H. *Proc. Natl. Acad. Sci. U.S.A,* **2005**, *102*, 5403.

[173] Liu, D.; Xu, Y.; Feng, Y.; Liu, H.; Shen, X.; Chen, K.; Ma, J.; Jiang, H. *Biochemistry,* **2006**, *45*, 10963.

Received: February 05, 2008 Revised: March 03, 2008 Accepted: March 05, 2008

PART III:
Modelling protein aggregation in bacteria

Available online at www.sciencedirect.com
SCIENCE @ DIRECT®

ELSEVIER

Journal of Biotechnology 125 (2006) 110–113

Journal of BIOTECHNOLOGY

www.elsevier.com/locate/jbiotec

Short communication

Protein activity in bacterial inclusion bodies correlates with predicted aggregation rates

Natalia Sánchez de Groot, Salvador Ventura*

*Departament de Bioquímica i Biologia Molecular, Institut de Biotecnologia i de Biomedicina,
Universitat Autònoma de Barcelona, E-08193 Bellaterra, Spain*

Received 14 December 2005; received in revised form 8 February 2006; accepted 14 February 2006

Abstract

Recent data show that protein aggregation as bacterial inclusion bodies does not necessarily imply loss of biological activity. Here, we investigate the effect of a large set of single-point mutants of an aggregation-prone protein on its specific activity once deposited in inclusion bodies. The activity of such aggregates significantly correlates with the predicted aggregation rates for each mutant, suggesting that rationally tuning the kinetic competition between folding and aggregation might result in highly active, inclusion bodies. The exploration of this technology during recombinant protein production would have a significant biotechnological value.

Keywords: Inclusion bodies; Recombinant protein expression; Protein aggregation; Protein folding; *Escherichia coli*

Protein misfolding is a common event during bacterial over-expression of recombinant genes (Baneyx and Mujacic, 2004). The aggregation of insoluble polypeptide chains as inclusion bodies (IBs) is the main bottleneck in protein production, narrowing the spectrum of relevant polypeptides obtained by recombinant techniques and hampering the development of top priority research areas such as the de novo design of novel proteins, the rational modification of natural proteins and structural genomics and proteomics. Being widespreadly believed that IB proteins are biologically inert and therefore useless in bioprocesses, many biologically relevant proteins have been disregarded for commercialisation.

We have shown recently that not only the aggregation of different recombinant proteins as bacterial IBs does not necessarily inactivate them but also that active IBs can be used in suspension as efficient catalysts for bioprocesses (Garcia-Fruitos et al., 2005). In concrete, the over-expression of a fusion of the aggregation-prone, Alzheimer-related peptide Aβ42 to green fluorescent protein (GFP) resulted in highly fluorescent IBs. In the present study, we have quantitatively investigated the biological activity of the IBs formed

* Corresponding author. Tel.: +34 93581 4147; fax: +34 93581 1264.
E-mail address: salvador.ventura@uab.es (S. Ventura).

0168-1656/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jbiotec.2006.02.026

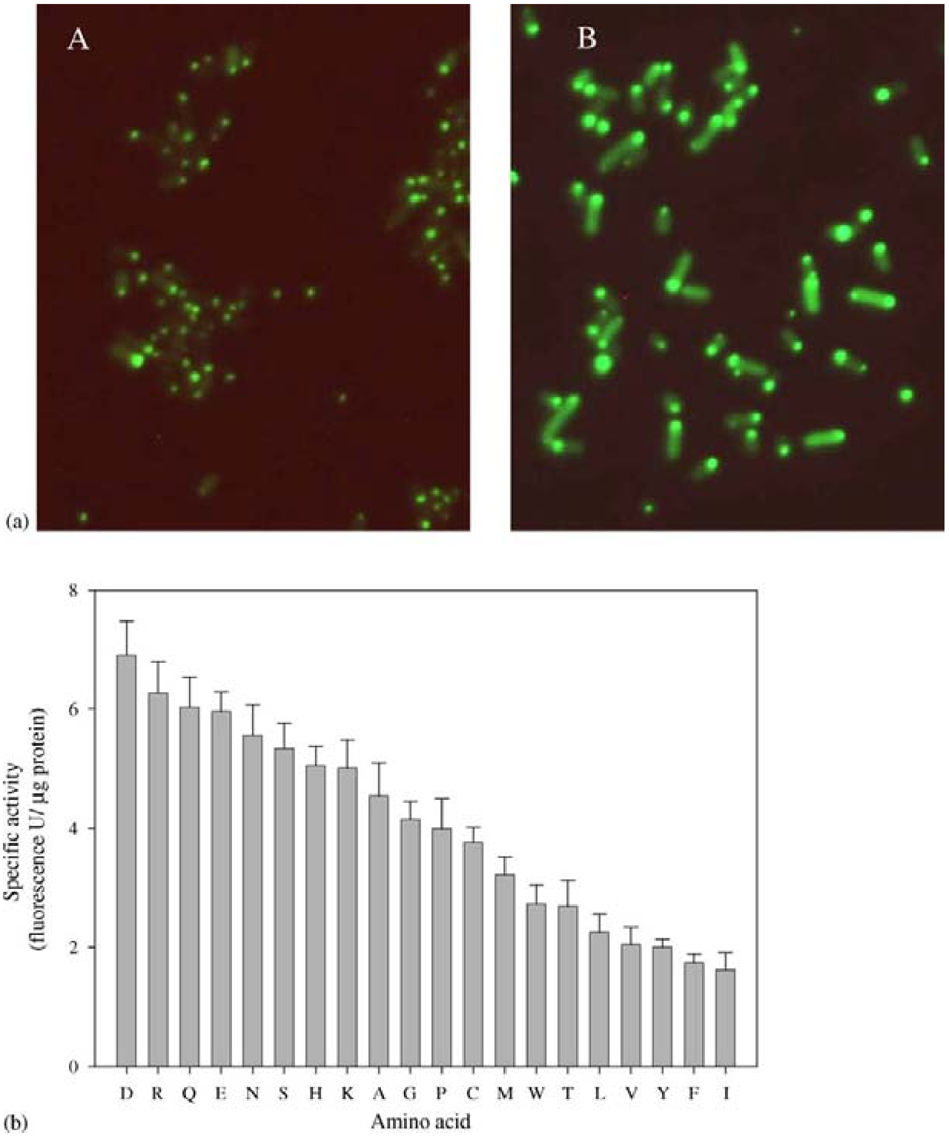


Fig. 1. Fluorescence emission of IBs formed by Aβ42-GFP variants (a) Fluorescence microscopy images of cells expressing wild type (A) and F19D (B) Aβ42-GFP. Cells were imaged at 40-fold magnification under UV light. Site-directed mutagenesis was performed using the QuickChange kit from Stratagene. All constructs were verified by DNA sequencing. Culture and protein expression conditions were as previously described (Garcia-Fruitos et al., 2005) with the exception that cultures were collected 8 h after induction to ensure equilibrium. (b) Specific activity of different Aβ42-GFP variants. The data are ordered by decreasing specific fluorescence at 510 nm (fluorescence U/µg protein). Inclusion bodies were purified by a detergent washing protocol as described (Carrio et al., 2000) and used in suspension for activity analysis. The GFP fluorescence of a 1 ml of suspension was recorded at 510 nm, using an excitation wavelength of 450 nm (emission and excitation slits widths 5 mm). Data were corrected for buffer signals. At least three different scans were averaged for each protein sample. For the determination of inclusion body protein, these structures were resuspended in denaturing buffer (Laemmli, 1970). After boiling for 20 min, appropriate sample volumes were loaded onto denaturing gels. Gels were scanned at high resolution and bands quantified by using the Quantity One software from Bio-Rad, by using appropriate protein dilutions of known concentration as controls. Determinations were always done within the linear range and they were used to calculate the specific activity values.

by 20 different mutants of the Aβ42-GFP fusion protein, to understand the rules underlying protein deposition during recombinant protein expression but also to explore the possibility of, through protein engineering, deliberately obtaining highly active proteins, useful for bioprocesses in IB form.

We used a set of 20 Aβ42-GFP fusion variants differing only in the residue at position 19 to elucidate if the primary sequence of a polypeptide might influence the occurrence of active protein in IBs when over-expressed in *Escherichia coli*. In wild type Aβ42 peptide the residue 19 is a Phe; each of the 19 mutants analyzed in this study possess a different natural amino acid in this position. Position 19 has been shown to strongly affect the aggregation of this Alzheimer-related peptide (Wood et al., 1995) thus being, a priori, a good target to test the effect of sequence changes on the specific activity of aggregated protein.

Upon overproduction, all 20 proteins formed fluorescent cytoplasmic inclusion bodies in *E. coli*. The different IBs were purified from the insoluble cell fraction and the intensity of the green fluorescence emitted by GFP embedded in IBs measured. The fluorescence emission changed from variant to variant (Fig. 1A). The specific fluorescence of the inclusion bodies formed by the most fluorescent mutant (F19D) was fourfold higher than those exhibited by the wild type aggregated protein (Fig. 1B).

In order to rationalize how sequence variation promotes changes in IBs activity we studied the correlation between IBs specific fluorescence and the predicted aggregation rates for the different Aβ42-GFP variants according to Eq. (1) (Fig. 2). A highly significant inverse correlation was observed ($r = 0.941$, $p = < 0.0001$), strongly suggesting that the final amount of active protein in a given IB depends on how fast the aggregation event occurs.

The results in this report indicate that the accumulation of active protein in IBs is not anecdotic but that it could be a general feature in recombinant protein production. More interestingly, we show that the aggregation of protein as IBs during recombinant protein expression is not an unspecific and passive process but rather a kinetically controlled event which speed depends specifically on the polypeptide nature and probably mainly on the sequence of certain aggregation-prone regions.

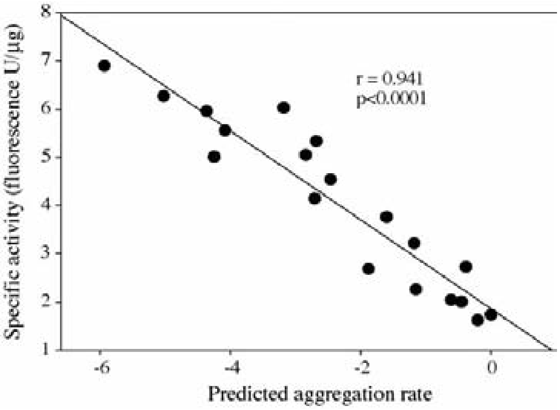


Fig. 2. Correlation between IBs specific fluorescence and predicted aggregation propensities of Aβ42-GFP fusions. Specific fluorescence of Aβ42-GFP fusions IBs were plotted vs. the changes in aggregation rates predicted by the Eq. (1) (developed by Chiti et al. (2003)). This approach assumes that β-sheet propensity, hydrophobicity and charge are independent factors, which affect the aggregation of a protein, in an additive manner.

$$\ln\left(\frac{v_{\mathrm{mut}}}{v_{\mathrm{wt}}}\right) = A\,\Delta\text{hydrophobicity} + B\,\Delta\beta\text{-sheet propensity} + C\,\Delta\text{charge} \qquad (1)$$

where v_{mut} and v_{wt} correspond to the predicted aggregation rates of the mutant and wild type sequences, respectively and Δcharge is the difference in the net charge of the polypeptide introduced by the mutation. *A*, *B* and *C* values are constants determined experimentally from the analysis of a large set of mutants of Acylphosphatase.

It is assumed, but scarcely proven in vivo, that aggregation competes with folding (Smith and Hall, 2001). The results herein constitute one of a few direct evidences of this theory. Fluorescence is indicative of both correct GFP folding and chromophore formation. Accordingly, the observed differences indicate different amount of active GFP trapped in the aggregates. Assuming that, in order to attain functionality, the attainment of a GFP native structure should precede aggregation, and that the time needed for this process is identical for GFP in all fusions assayed, then IBs fluorescence emission relates to the time the Aβ42-GFP variant was soluble after its synthesis and before to its aggregation. Thus, fluorescence probably inversely reflects the in vivo aggregation rate as suggested by the highly significant correlation found between IBs fluorescence and predicted aggregation rates. The faster the fusion protein aggregates, the lower its fluorescence emission and vice versa, in such a way that fluorescent molecules in IBs are those whose aggregation was slow enough to permit prior GFP folding. It is important to note that usually the productive folding of the GFP reporter has been directly related to the solubility of the upstream fused protein when over-expressed in *E. coli* (Waldo et al., 1999). According to our data it could have eventually been indicative of the folding performance of the fused protein rather than of its solubility–insolubility.

Overall, this system appears as a valid one to explore in vivo the kinetic competition between folding and aggregation events. As shown here, tuning aggregation speed might result in highly active (fluorescent) IBs, which being highly pure, compact but porous and hydrated protein microparticles might be used as bio catalysers opening intriguing possibilities for the biotechnological industry.

Acknowledgments

We thank Prof. Antonio Villaverde for manuscript revision and suggestions. We also thank Prof. Josep Vendrell and Prof. Francesc Xavier Aviles for lab facilities. S.V. is recipient of a “Ramón y Cajal” contract awarded by the MCYT-Spain and co-financed by the Universitat Autònoma de Barcelona (UAB, http://www.uab.es/), and founded by PNL2004-40 (UAB) and 2005SGR-00037(AGAUR).

References

Baneyx, F., Mujacic, M., 2004. Recombinant protein folding and misfolding in *Escherichia coli*. Nat. Biotechnol. 22, 1399–1408.

Carrio, M.M., Cubarsi, R., Villaverde, A., 2000. Fine architecture of bacterial inclusion bodies. FEBS Lett. 471, 7–11.

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.

Garcia-Fruitos, E., Gonzalez-Montalban, N., Morell, M., Vera, A., Ferraz, R.M., Aris, A., Ventura, S., Villaverde, A., 2005. Aggregation as bacterial inclusion bodies does not imply inactivation of enzymes and fluorescent proteins. Microb. Cell Factors 4, 27.

Laemmli, U.K., 1970. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature 227, 680–685.

Smith, A.V., Hall, C.K., 2001. Protein refolding versus aggregation: computer simulations on an intermediate-resolution protein model. J. Mol. Biol. 312, 187–202.

Waldo, G.S., Standish, B.M., Berendzen, J., Terwilliger, T.C., 1999. Rapid protein-folding assay using green fluorescent protein. Nat. Biotechnol. 17, 691–695.

Wood, S.J., Wetzel, R., Martin, J.D., Hurle, M.R., 1995. Prolines and amyloidogenicity in fragments of the Alzheimer’s peptide AbetaA42. Biochemistry 34, 724–730.

Effect of temperature on protein quality in bacterial inclusion bodies

Natalia Sánchez de Groot[a], Salvador Ventura[a,b,*]

[a] *Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, E-08193 Bellaterra, Spain*
[b] *Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bioquimica, E-08193 Bellaterra, Spain*

Received 27 September 2006; revised 17 October 2006; accepted 31 October 2006

Available online 10 November 2006

Edited by Judit Ovádi

Abstract Increasing evidence indicates that protein aggregation in bacteria does not necessarily imply loss of biological activity. Here, we have investigated the effect of growth-temperature on both the activity and stability of the inclusion bodies formed by a point-mutant of Aβ42 Alzheimer peptide, using green fluorescent protein as a reporter. The activity in the aggregates inversely correlates with the temperature. In contrast, inclusion bodies become more stable in front of chemical denaturation and proteolysis when temperature increases. Overall, the data herein open new perspectives in protein production, while suggesting a kinetic competition between protein folding and aggregation during recombinant protein expression.

Keywords: Inclusion bodies; Protein aggregation; Protein folding; Recombinant protein expression; *Escherichia coli*

1. Introduction

In many cases, the production of recombinant polypeptides in prokaryotic hosts results in incomplete folding processes that usually end with their accumulation as insoluble aggregates, known as inclusion bodies (IBs), in the cytoplasm and/or in the periplasmic space of the cells [1–3]. The aggregation of insoluble polypeptide chains as IBs is of major concern in biotechnology, since it prevents the commercialisation of many relevant polypeptides [3]. The complete aggregation process is still poorly understood. The current view about IBs has recently evolved from considering proteins in IBs as totally inactive to accept that aggregation of recombinant proteins as bacterial IBs does not necessarily inactivate them [4–7]. This allows using the activity of the protein embedded in IBs as a reporter to monitor the influence of both intrinsic and extrinsic factors on the aggregation process. This could become an important subject in biotechnology since it can be of help in tuning optimal sequential and culture conditions for protein production.

We have shown that, the overexpression of a fusion of the aggregation-prone, Alzheimer-related, peptide Aβ42 to green fluorescent protein (GFP) results in fluorescent IBs. Using this system we have investigated the effect of the polypeptide sequence on protein quality in bacterial aggregates [8]. The approach also permits to easily monitor the effects of extrinsic factors, such as growth temperature, on bacterial protein aggregation.

A widespread strategy to reduce the in vivo aggregation of recombinant polypeptides consists of cultivation at reduced temperatures [9]. This approach has proven effective in increasing the solubility of a number of difficult proteins at expenses of the final yield [10]. However, only recently it has been addressed the effect of the temperature on the characteristics of the aggregated fraction [5,11–13]. In the present investigation, we have quantitatively investigated the biological activity and the stability of the IBs formed by a variant of the Aβ42-GFP fusion protein at different cultivation temperatures to provide insights into the rules and polypeptide interactions underlying protein deposition during recombinant protein expression.

2. Material and methods

2.1. Protein expression and IBs purification

Escherichia coli BL21(DE3) was used for all the experiments. Plasmid encoding Aβ42(F19D)-GFP has been previously described [8]. Cells expressing the Aβ42-GFP fusion were grown for 4 h at 37 °C in LB medium containing 35 μg/ml kanamycin and pre-incubated at the selected expression temperature for 30 min. Then, protein expression was induced with 1 mM IPTG. Cultures were grown at the selected temperatures for 20 additional hours, to ensure fluorescence equilibrium, and harvested by centrifugation. Expression of Aβ-GFP fusion protein was monitored by SDS–PAGE using a 12% (w/v) gel. Inclusion bodies (IBs) were purified from cell extracts by detergent-based procedures as described [14]. For the determination of inclusion body protein, these structures were resuspended in denaturing buffer [15]. After boiling for 10 min, appropriate sample volumes were loaded onto denaturing gels. Gels were scanned at high resolution and bands quantified by using the Quantity One software from Bio Rad, by employing appropriate protein dilutions of known concentration as controls. Determinations were always done within the linear range and they were used to calculate the specific activity values.

2.2. Fluorescence measurements

Emission spectra of GFP in IBs were measured on a Perkin–Elmer 650-40 spectrofluorimeter (Boston, MA, USA). The GFP fluorescence of a 1 ml of IBs suspension in 10 mM Tris–HCl buffer (pH 7.5) was recorded from 500 to 600 nm, using an excitation wavelength of 470 nm. Emission and excitation slits width were fixed at 10 and 5 mm, respectively. Dilutions were employed when necessary and data were corrected for buffer signals and protein concentration. At least three different scans were averaged for each IBs sample. For microscopy analysis, IBs formed at different temperatures were isolated from the insoluble cell fraction by repeated detergent washing as described [14] and deposited on top of glass slides. Images of purified IBs were obtained at 40-fold magnification under UV light or using phase contrast in a Leica DMBR microscope. The average size of purified IBs was measured under phase contrast by analysing 40 individual aggregates corresponding to two different fields for each temperature, using the Leica QWin Standard V2.3 software.

*Corresponding author. Fax: +34 93 581 12 64.
E-mail address: salvador.ventura@uab.es (S. Ventura).

0014-5793/$32.00 © 2006 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2006.10.071

2.3. Proteolytic digestion of IBs

Purified Aβ42(F19D)-GFP IBs, obtained at different temperatures, were passed 10 times through a 0,1 mm needle to homogenize the aggregate solutions. The IBs were diluted at 1 $OD_{350\,nm}$ in 792 μl of 50 mM Tris–HCl, 150 mM NaCl buffer (pH 8.0). 8 μl of concentrated proteinase K was added the IBs solution to obtain a 0.2 mg/ml final concentration and initiate the proteolytic reaction. The digestion was monitored for 150 min by measuring the changes in $OD_{350\,nm}$ in a Cary-100 Varian spectrophotometer.

2.4. Stability of IBs in front of chemical solubilization

50 μl of a 1 $OD_{350\,nm}$ solution of purified and homogenized Aβ42(F19D)-GFP IBs, obtained at different temperatures, was added to 950 μl of 10 mM Tris–HCl buffer (pH 7.5) containing selected concentrations of guanidinium hydrochloride (ranging from 0 to 6 M) for equilibrium denaturation experiments. The reactions were allowed to reach equilibrium by incubating them for 20 h at room temperature. The effect of the denaturant was measured by monitoring the changes in $OD_{350\,nm}$ in a Cary-100 Varian spectrophotometer. The fitting of the experimental data was performed using the non-linear, least-squares algorithm provided with the software KaleidaGraph (Abelbeck Software) assuming a two-state solubilization mechanism.

For kinetic experiments, 50 μl of a 1 $OD_{350\,nm}$ solution of purified and homogenized Aβ42(F19D)-GFP IBs, obtained at different temperatures, was added to 950 of 10 mM Tris–HCl buffer (pH 7.5) containing 2,3 M guanidinium hydrochloride. The reaction was monitored for 180 min by measuring the changes in $OD_{350\,nm}$ in a Cary-100 Varian spectrophotometer. Double-exponential decay curves were fitted to the data using Sigmaplot non-linear regression software (Jandel Scientific, San Rafael, CA, USA), and apparent rate constants were derived from these regressions.

3. Results and discussion

3.1. Effect of the growth temperature on the activity of IBs

In a previous study we have used the Alzheimer related Aβ42 gene fused upstream of the GFP sequence and under the control of the T7 promoter as a model to investigate aggregation in bacteria. At 37 °C, *E. coli* cells transformed with this construction express, upon IPTG induction, a high amount of the Aβ42-GFP fusion protein that accumulates in active fluorescent cytoplasmatic inclusion bodies [8]. We have shown that mutation of Phe in position 19 of Aβ42 to Asp abolish the amyloidogenicity of this Alzheimer-related peptide [16]. This change also promotes a fourfold increase in the specific fluorescence emitted by IBs relative to that emitted by the wild type Aβ42 when fused to GFP and expressed at 37 °C in *E. coli* [8]. This mutant provides a wider dynamic range to explore the effects of extrinsic factors on the fluorescence of IBs, especially if conditions expected to decrease the activity of these aggregates are going to be tested. In the present study we explored whether the temperature of cultivation influences the activity of the protein embedded in the aggregates or if on the contrary the fluorescence of the IBs is independent of the temperature at which they are formed. To this aim we expressed the Aβ42(F19D)-GFP fusion at temperatures ranging from 18 to 42 °C. A fraction of the expressed protein fusion accumulated as insoluble IBs at all the temperatures assayed. We purified the different inclusion bodies and compared their specific activity by measuring GFP fluorescence emission. As shown in Fig. 1 the activity of the protein in IBs is strongly influenced by the temperature of cultivation. Increasing the growth temperature above 37 °C results in reduction in specific activity, while lowering it significantly increases the fluorescence emission. The specific fluorescence of the IBs formed at 18 °C was 16-fold higher than those exhibited by the IBs purified from cells cultured at 42 °C. The influence of temperature on IBs fluorescence emission, once purified from the insoluble cell fraction, can be also visually analyzed by using fluorescence microscopy (Fig. 2). In agreement with the spectral data, IBs formed at low temperatures appear clearly as a more fluorescent aggregates than those formed at high temperature. Using phase contrast it could be observed that the isolated aggregates formed at 42 °C, 37 °C, 30 °C, and 25 °C all display similar sizes, ranging from 1.1 to 1.2 μm. The IBs formed at 18 °C appear as smaller, much less refractile aggregates with an average size of about 0.9 μm. Interestingly, similar morphological and fluorescent properties have been reported recently for the aggregates formed by a VP1-GFP fusion at 16 °C [12]. In order to rationalize how temperature promotes changes in IBs activity, we studied the correlation between IBs specific fluorescence and the cultivation temperature. A strong inverse correlation was observed indicating a linear dependence of IBs activity on the growth temperature in the studied range (Fig. 1B) and thus an increase in the proportion of native-like conformations in IBs formed at low temperatures.

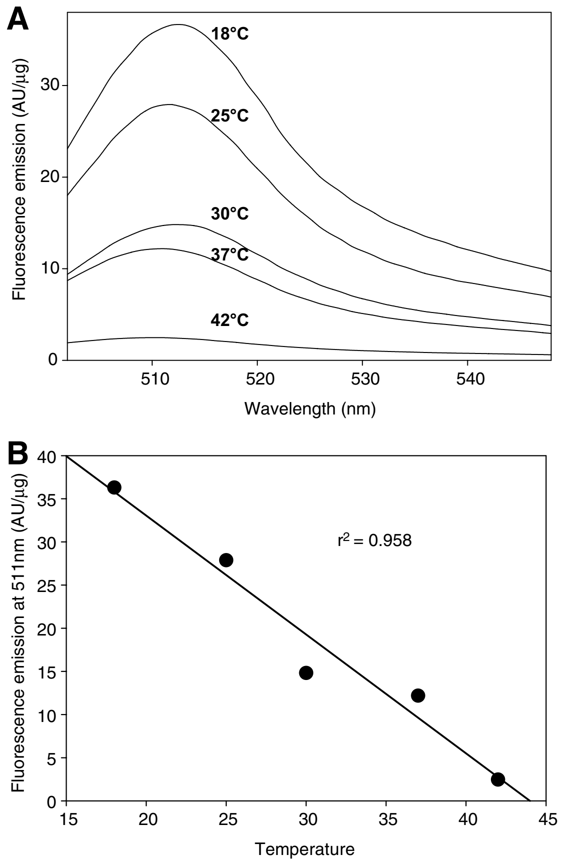


Fig. 1. Dependence of the specific fluorescence emission of Aβ42(F19D)-GFP IBs on the growth temperature. (A) Fluorescence spectra of GFP in IBs at selected temperatures. (B) Correlation between IBs activity and temperature of cultivation.

3.2. Effect of the growth temperature on the stability of IBs

It is thought that during recombinant protein production aggregation is in general favoured at higher temperatures due to the strong temperature dependence of hydrophobic

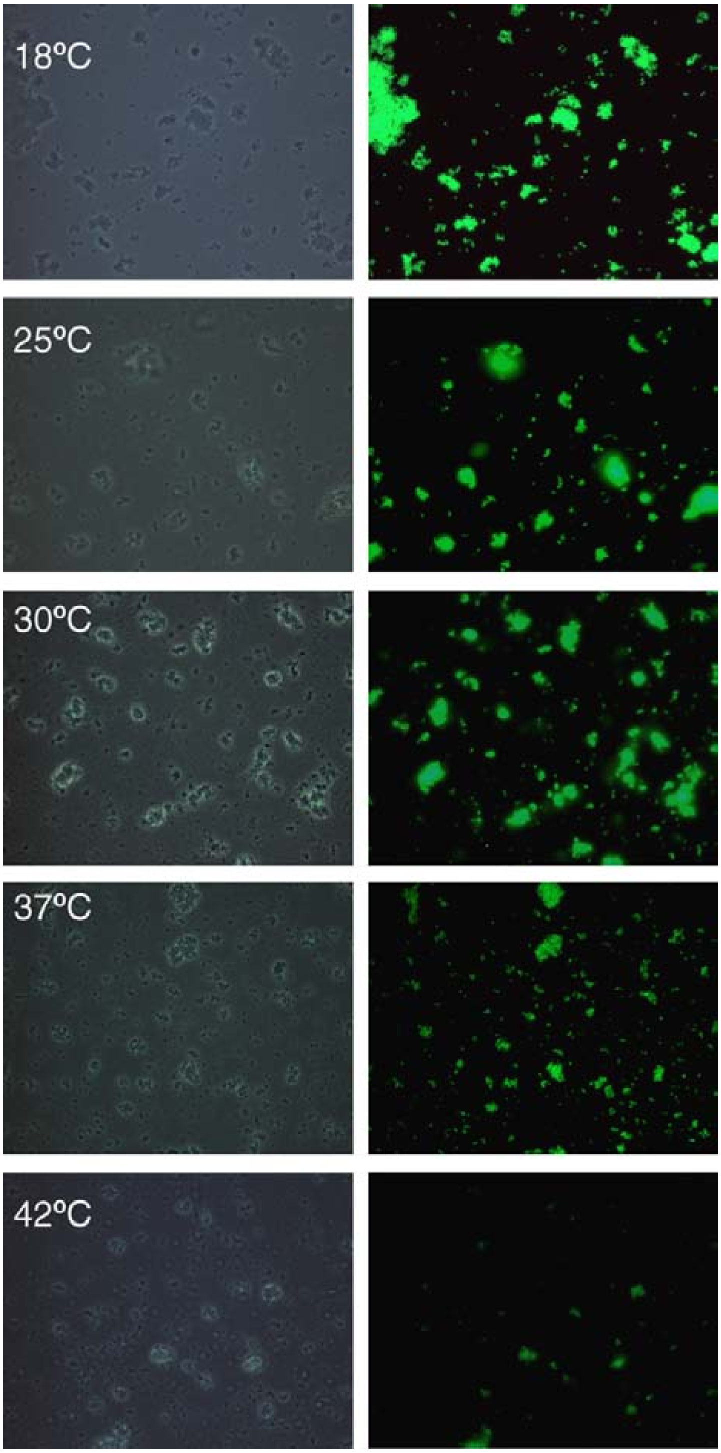


Fig. 2. Visualization of GFP fluorescence in isolated IBs formed at different temperatures. The left series correspond to phase contrast microscopy of purified IBs, the right series to fluorescence microscopy under UV light, both with 40-fold magnification.

interactions involved in the aggregation reaction [17,18]. Nevertheless, and to the best of our knowledge, it has not been investigated yet whether this results in a dependence of IBs conformational stability on the growth temperature. To explore this possibility, we compared first the resistance to proteinase K digestion of the IBs formed under standard conditions (37 °C) with that of IBs formed at higher (42 °C), and lower (25 °C) temperatures. Proteinase K is an endolytic serine protease that cleaves peptide bonds at the carboxylic sides of aliphatic, aromatic or hydrophobic amino acids. It has find application in the mapping of polypeptide regions in the core of amyloid fibrils due to its strong preference for hydrolyzing unstructured protein regions [19]. We monitored the kinetics of IBs sensitivity to proteolysis by measuring the decrease in turbidity at 350 nm upon addition of proteinase K. As illustrated in Fig. 3, IBs formed at high temperature

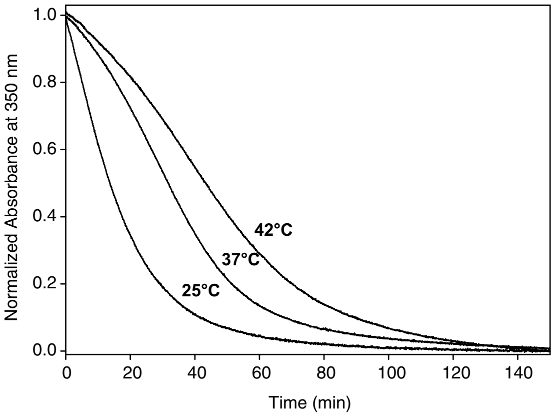


Fig. 3. Kinetics of IBs' proteolytic digestion monitored by a time-dependent decrease of turbidity at 350 nm. The growth temperatures at which inclusion bodies were obtained are indicated on top of the curves.

are clearly more stable in front of proteolysis than those formed at low temperature. The IBs formed at 42 °C and 37 °C exhibited a sigmoid curve that suggests an initial higher resistance to digestion, and thus a more densely packed structure in the initial aggregated species that is lost after the initial proteolytic attack. This effect is not appreciable in the IBs formed at 25 °C, indicating a high sensibility to protease action already at the beginning of the experiment, probably due to the presence of a higher amount of accessible polypeptide chains.

We investigated if the differential stability of these IBs in front of proteases correlates with their resistance in front of chemical denaturation. To this aim, we dissolved the different IBs in buffer containing selected concentrations of guanidinium hydrochloride (Gnd HCl). This chaotropic agent has been used recently to study the resistance to solubilization of the IBs and thermal aggregates formed by different proteins [20]. The reaction was typically performed at room temperature for 20 h to allow equilibrium; then, the effect of the denaturant was measured by monitoring the changes in absorbance at 350 nm. We assumed a two-state mechanism in which the protein is either in an aggregated state that contributes to turbidity or in a soluble state which does not contributes to the absorbance at 350 nm (independently of the fact that the protein could be properly folded or not in the aggregated or soluble states). Although this assumption is clearly a simplification of the effect of the chaotropic agent on IBs, the curves could be properly fitted to a two states process ($R = 0.999$ in all cases) (Fig. 4). From the data it can be clearly inferred that IBs formed at different temperatures differ also in their conformational stability against chemical denaturation, being again the IBs formed at 42 °C (transition midpoint = 1.72 M Gnd HCl) more tolerant to the presence of Gnd HCl than those formed at 37 °C (transition midpoint = 1,45 M Gnd HCl) or 25 °C (transition midpoint = 1,36 M Gnd HCl). To further confirm this point we sought to analyze the kinetics of solubilization of the three different IBs by a fixed concentration of denaturant. This way, we incubated the IBs formed at 42 °C, 37 °C and 25 °C in 2,3 M Gnd HCl and monitored the dependence of the turbidity signal at 350 nm on the time (Fig. 5). The data could be fitted to a double-exponential decay equation with very good accuracy ($R > 0.99$) and the differences in the apparent rate constants of the fast phase calculated. Significant differences in the velocity of solubilization could be observed between samples, with 0.376 ± 0.005, 0.290 ± 0.002 and 0.219 ± 0.002 min^{-1} fast rate constants for IBs formed at 25 °C, 37 °C and 42 °C, respectively. Thus, in excellent agreement with the equilibrium data, the IBs formed at low temperatures are solubilized faster than those formed at higher temperatures, indicating that the cultivation temperature determines the stability, and thus the conformational properties of the polypeptide chains embedded in bacterial aggregates.

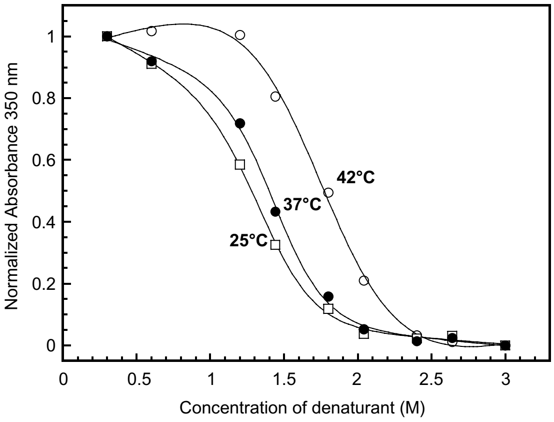


Fig. 4. Stability of IBs in front of Gnd HCl denaturation under equilibrium conditions. The growth temperatures at which inclusion bodies were obtained are indicated on side of the curves.

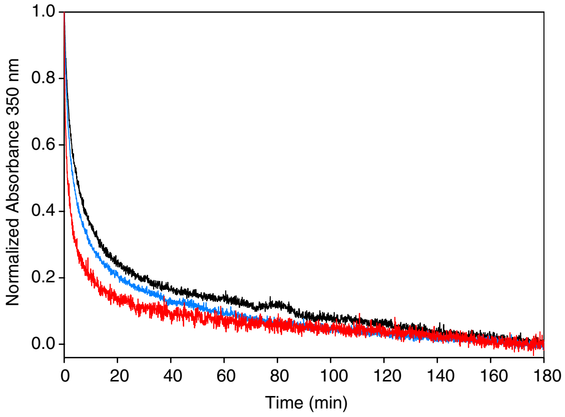


Fig. 5. Kinetics of solubilization by 2,3 M Gnd HCl of IBs formed at 42 °C (black), 37 °C (blue) and 25 °C (red), monitored by a time-dependent decrease of turbidity at 350 nm.

3.3. Relationship between IBs conformational stability and activity

To decipher if the solubilization of IBs by chemical denaturation affects the activity of the protein embedded in these bacterial aggregates, we investigated the effect of Gnd HCl on IBs activity and compared it with the impact on IBs conformational stability. We proceeded as described for the equilibrium denaturation experiment, but this time we monitored the dependence of GFP fluorescence emission on denaturant concentration. In Fig. 6 the equilibrium curves obtained by monitoring absorbance at 350 nm and protein fluorescence in IBs

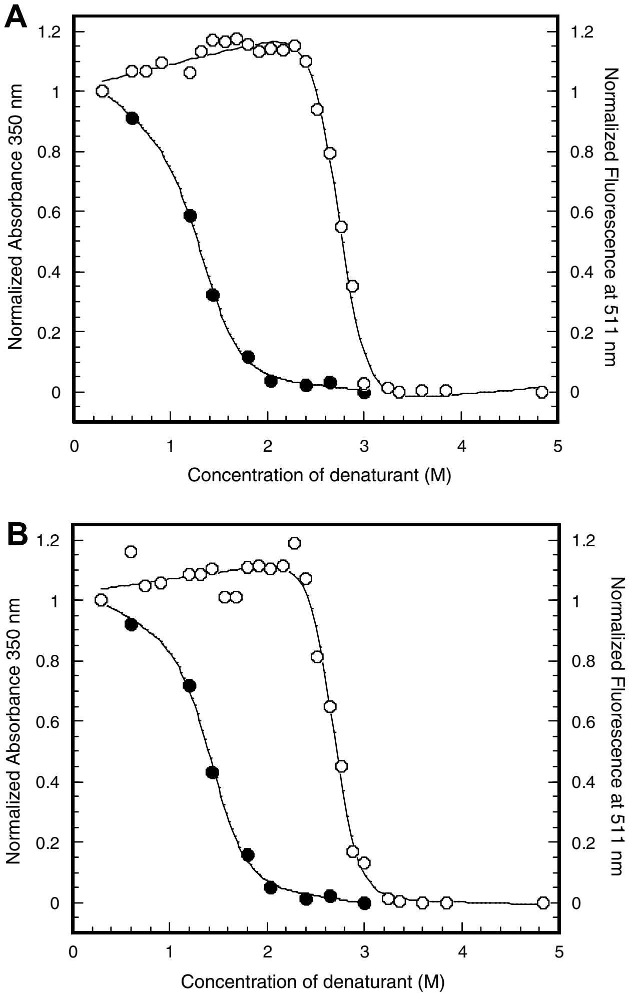


Fig. 6. Equilibrium dependence of the turbidity signal (solid symbols) and GFP fluorescence (empty symbols) of IBs formed at 25 °C (A) and 37 °C (B) on the Gnd HCl concentration.

growth at 25 and 37 °C are compared (the low activity of IBs formed at 42 °C prevented to record accurate data at high denaturant concentrations). Surprisingly, it is observed that the turbidity signal disappears at chaotropic agent concentrations in which GFP is still fully active in both types of IBs, indicating that the turbidity curve provides us mainly information about the loss of the intermolecular interactions that stabilize IBs rather than on intramolecular contacts, which would account for the native conformation and activity of GFP. This results has important implications for recombinant protein production since suggests that proteins can be liberated from IBs in a fully functional state, by using conditions which specifically disturb the network of intermolecular contacts that provides stability to IBs without denaturating the native protein embedded in the aggregates. These data are in agreement with the suggestion that the presence of native-like structures within IBs could improve the efficiency of refolding strategies that use mild solubilization conditions [21], as well as with the observation that the in vitro refolding of IBs formed at low temperature renders higher yields of active polypeptides than when employing IBs constructed at higher temperatures [22].

From the present data, it follows that high temperatures promote stable aggregates because they favour intermolecular interactions at expenses of native intramolecular contacts and thus protein activity. This way the lower activity of IBs produced at high temperature indicates a higher proportion of non-native protein conformations respect to IBs formed at low temperature. In principle, this non-properly folded polypeptide chains or segments are ready to establish intermolecular interactions among them in the aggregates promoting their stability. These contacts are more likely involved in the formation and stabilization of the intermolecular β-sheet structure recently described to be common to IBs formed by unrelated proteins [23–25]. On the contrary, production of protein at low temperatures results in highly active IBs which indicates that in this molecules, aggregation-promoting regions are likely to be blocked in the native state of globular GFP because their side chains are hidden in the inner hydrophobic core or already involved in the network of contacts that stabilizes the native state of a protein [26]. The lower number of unfolded, aggregation-prone chains available to establish the intermolecular interactions that glue the structure of IBs would explain the lower conformational stability of these low temperature aggregates. It is also deduced that, at least in this particular case, the intramolecular native contacts that maintain the folded protein structure are stronger than non-native interactions between polypeptide chains, since they resist clearly higher denaturant concentrations. Interestingly, Villaverde and co-workers have recently reported, using GFP and a VP1-GFP fusion, that, in excellent agreement with the present data, low temperature cultivation results in more active IBs [12]. Although the stability of the different IBs was not addressed in this work, it was demonstrated that the intermolecular extended β-sheet conformation of IBs loosed compactness at lower temperatures as demonstrated by ATR-FTIR. Similar structural results have been found for recombinant growth hormone, human interferon-alpha-2b and a lipase when expressed in *E. coli* as IBs [5,27]. In these particular cases, the relative intensity between the native and the aggregated IR contributions was shown to be modulated by protein expression levels. Overall, this conjunct of recently collected data provides evidence that during protein recombinant production the processes of folding to attain a native conformation stabilized by intrachain contacts and aggregation to attain a non-native structure stabilized by interchain interactions are competing. As shown here, for a given polypeptide, the equilibrium can be shifted in either direction by specific extrinsic factors.

An increasing body of evidence [1,28,29] suggests that the possibility to find out strategies that favour in vivo folding versus aggregation would open intriguing opportunities both in the protein production and protein aggregation research. This way, the production in bacteria of highly active and poorly stable IBs by modulating the culture conditions, together with the development of simple and economic downstream strategies, such us mild IBs disaggregation (without the requirement for aggressive unfolding/refolding steps) appears as a promising avenue for the production of difficult proteins, such us mammalian ones, in a soluble, properly folded and active conformation useful for biotechnological applications.

Acknowledgements: We thank Prof. Josep Vendrell and Prof. Francesc Xavier Aviles for lab facilities and helpful discussions. We thank

Amanda Vidaña for technical assistance. This work was supported by (BIO2004-05879) (Ministerio de Ciencia y Tecnología-Spain) and 2005SGR-00037(AGAUR). N.S. is recipient of a Ministerio Educacion y Ciencia-Spain fellowship.

References

[1] Baneyx, F. and Mujacic, M. (2004) Recombinant protein folding and misfolding in *Escherichia coli*. Nat. Biotechnol. 22, 1399–1408.

[2] Villaverde, A. and Carrio, M.M. (2003) Protein aggregation in recombinant bacteria: biological role of inclusion bodies. Biotechnol. Lett. 25, 1385–1395.

[3] Fahnert, B., Lilie, H. and Neubauer, P. (2004) Inclusion bodies: formation and utilisation. Adv. Biochem. Eng. Biotechnol. 89, 93–142.

[4] Garcia-Fruitos, E., Gonzalez-Montalban, N., Morell, M., Vera, A., Ferraz, R., Aris, A., Ventura, S. and Villaverde, A. (2005) Aggregation as bacterial inclusion bodies does not imply inactivation of enzymes and fluorescent proteins. Microbial Cell Factories 4, 27.

[5] Ami, D., Natalello, A., Gatti-Lafranconi, P., Lotti, M. and Doglia, S.M. (2005) Kinetics of inclusion body formation studied in intact cells by FT-IR spectroscopy. FEBS Lett. 579, 3433–3436.

[6] Ventura, S. and Villaverde, A. (2006) Protein quality in bacterial inclusion bodies. Trends Biotechnol. 24, 179–185.

[7] Tokatlidis, K., Dhurjati, P., Millet, J., Beguin, P. and Aubert, J.P. (1991) High activity of inclusion bodies formed in *Escherichia coli* overproducing *Clostridium thermocellum* endoglucanase D. FEBS Lett. 282, 205–208.

[8] de Groot, N.S. and Ventura, S. (2006) Protein activity in bacterial inclusion bodies correlates with predicted aggregation rates. J. Biotechnol. 125, 110–113.

[9] Sorensen, H.P. and Mortensen, K.K. (2005) Soluble expression of recombinant proteins in the cytoplasm of *Escherichia coli*. Microbial Cell Factories 4, 1.

[10] Vasina, J.A. and Baneyx, F. (1996) Recombinant protein expression at low temperatures under the transcriptional control of the major *Escherichia coli* cold shock promoter cspA. Appl. Environ. Microbiol. 62, 1444–1447.

[11] Schultz, T., Martinez, L. and de Marco, A. (2006) The evaluation of the factors that cause aggregation during recombinant expression in *E. coli* is simplified by the employment of an aggregation-sensitive reporter. Microbial Cell Factories 5, 28.

[12] Vera, A., Gonzalez-Montalban, N., Aris, A. and Villaverde, A. (2006) The conformational quality of insoluble recombinant proteins is enhanced at low growth temperatures. Biotechnol. Bioeng.

[13] Hunke, S. and Betton, J.M. (2003) Temperature effect on inclusion body formation and stress response in the periplasm of *Escherichia coli*. Mol. Microbiol. 50, 1579–1589.

[14] Carrio, M.M., Cubarsi, R. and Villaverde, A. (2000) Fine architecture of bacterial inclusion bodies. FEBS Lett. 471, 7–11.

[15] Laemmli, U.K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature 227, 680–685.

[16] de Groot, N.S., Aviles, F.X., Vendrell, J. and Ventura, S. (2006) Mutagenesis of the central hydrophobic cluster in Abeta42 Alzheimer's peptide. Side-chain properties correlate with aggregation propensities. FEBS J. 273, 658–668.

[17] Kiefhaber, T., Rudolph, R., Kohler, H.H. and Buchner, J. (1991) Protein aggregation in vitro and in vivo: a quantitative model of the kinetic competition between folding and aggregation. Biotechnology (N Y) 9, 825–829.

[18] Schellman, J.A. (1997) Temperature, stability, and the hydrophobic interaction. Biophys. J. 73, 2960–2964.

[19] Bocharova, O.V., Makarava, N., Breydo, L., Anderson, M., Salnikov, V.V. and Baskakov, I.V. (2006) Annealing prion protein amyloid fibrils at high temperature results in extension of a proteinase K-resistant core. J. Biol. Chem. 281, 2373–2379.

[20] Rinas, U., Hoffmann, F., Betiku, E., Estape, D. and Marten, S. (2006) Inclusion body anatomy and functioning of chaperone-mediated in vivo inclusion body disassembly during high-level recombinant protein production in *Escherichia coli*. J. Biotechnol.

[21] Patra, A.K., Mukhopadhyay, R., Mukhija, R., Krishnan, A., Garg, L.C. and Panda, A.K. (2000) Optimization of inclusion body solubilization and renaturation of recombinant human growth hormone from *Escherichia coli*. Protein Expr. Purif. 18, 182–192.

[22] Jevsevar, S., Gaberc-Porekar, V., Fonda, I., Podobnik, B., Grdadolnik, J. and Menart, V. (2005) Production of nonclassical inclusion bodies from which correctly folded protein can be extracted. Biotechnol. Prog. 21, 632–639.

[23] Przybycien, T.M., Dunn, J.P., Valax, P. and Georgiou, G. (1994) Secondary structure characterization of beta-lactamase inclusion bodies. Protein Eng. 7, 131–136.

[24] Ami, D., Bonecchi, L., Cali, S., Orsini, G., Tonon, G. and Doglia, S.M. (2003) FT-IR study of heterologous protein expression in recombinant *Escherichia coli* strains. Biochim. Biophys. Acta 1624, 6–10.

[25] Carrio, M., Gonzalez-Montalban, N., Vera, A., Villaverde, A. and Ventura, S. (2005) Amyloid-like properties of bacterial inclusion bodies. J. Mol. Biol. 347, 1025–1037.

[26] Ventura, S. (2005) Sequence determinants of protein aggregation: tools to increase protein solubility. Microbial Cell Factories 4, 11.

[27] Ami, D., Natalello, A., Taylor, G., Tonon, G. and Maria Doglia, S. (2006) Structural analysis of protein inclusion bodies by Fourier transform infrared microspectroscopy. Biochim. Biophys. Acta 1764, 793–799.

[28] Singh, S.M. and Panda, A.K. (2005) Solubilization and refolding of bacterial inclusion body proteins. J. Biosci. Bioeng. 99, 303–310.

[29] Schrodel, A. and de Marco, A. (2005) Characterization of the aggregates formed during recombinant protein expression in bacteria. BMC Biochem. 6, 10.

Studies on bacterial inclusion bodies

Natalia S de Groot,
Alba Espargaró,
Montserrat Morell &
Salvador Ventura[†]

[†]*Author for correspondence*
Departament de Bioquímica i Biologia Molecular &
Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona,
E-08193 Bellaterra, Spain
Tel.: +34 935 868 147;
Fax: +34 935 811 264;
salvador.ventura@uab.es

The field of protein misfolding and aggregation has become an extremely active area of research in recent years. Of particular interest is the deposition of polypeptides into inclusion bodies inside bacterial cells. One reason for this interest is that protein aggregation constitutes a major bottleneck in protein production and restricts the spectrum of protein-based drugs available for commercialization. Additionally, prokaryotic cells could provide a simple yet powerful system for studying the formation and prevention of toxic aggregates, such as those responsible for a number of degenerative diseases. Here, we review recent work that has challenged our understanding of the structure and physiology of inclusion bodies and provided us with a new view of intracellular protein deposition, which has important implications in microbiology, biomedicine and biotechnology.

In the last few years, protein aggregation has evolved from a neglected area of protein chemistry to become an important subject in many fields, including biology, medicine and biotechnology [1]. An increasing body of evidence has shown that the anomalous disassembly of proteins is the fundamental cause behind certain debilitating human diseases of growing incidence, such as Alzheimer's, Parkinson's, type II diabetes and many others [2–4]. It has been assumed that the formation of insoluble aggregates was directly linked to the onset of these pathologies. However, recent data suggest that pre-aggregated, diffusible assemblies are the most harmful species and that aggregates might, in fact, have a protective role [5]. Additionally, the aggregation of proteins in prokaryotic cells such as insoluble inclusion bodies (IBs) is a major bottleneck in the protein production pipeline, narrowing the spectrum of polypeptides that are available for priority research areas, such as structural genomics or proteomics. It also has a huge economic impact on the biotechnology market, preventing the production of many relevant protein-based drugs.

Bacteria, specifically *Escherichia coli*, are widely used as factories for the production of recombinant polypeptides that do not require post-translational modifications to achieve their native and bioactive conformations. *E. coli* grows to high cell density rapidly with inexpensive substrates and offers inducible protein expression at extremely high levels. However, its intrinsic propensity to accumulate heterologous products in IBs presents a major challenge for downstream bioprocessing. Recovering the target protein in an active conformation from these insoluble deposits through successive unfolding and refolding steps is cumbersome; usually it results in low recovery and significantly increases production time and cost.

For a long time, biotechnological efforts to improve protein production have been focused on increasing protein solubility and reducing IB formation. Nevertheless, the formation of IBs is often unavoidable. In some large structural genomic projects, up to half of the targets tested failed to fold properly and instead accumulated as insoluble protein [6]. In contrast to the large effort devoted to the study of amyloid aggregates related to conformational diseases, traditionally little attention has been paid to the structural and functional characteristics of these intracellular aggregates in bacteria. Recent studies demonstrate, however, that protein aggregation in bacterial IBs and amyloid fibrils share several traits. Therefore, bacterial systems should be able to provide new insights into structural and/or sequential constraints underlying protein deposition in a biologically relevant context; this could be helpful for developing new drugs and therapies. In the last few years, this potential has driven the collection of new, relevant data on the physiology and structure of IBs, as well as the dominant forces causing aggregation into IBs. This new work is opening an avenue for the development of an integrated model of intracellular protein aggregation in bacteria. In addition, the emerging information has encouraged the development of new strategies to increase protein productivity and quality in biotechnological processes.

Keywords: amyloid, chaperone, *Escherichia coli*, inclusion bodies, protein aggregation, protein folding, protein production, protein solubility

future medicine part of fsg

10.2217/17460913.3.4.423 © 2008 Future Medicine Ltd ISSN 1746-0913

What is an IB?

Anatomy & composition

IBs are insoluble protein aggregates that are frequently observed in bacteria following overexpression of heterologous genes whose products fail to attain a soluble, bioactive conformation. Using phase contrast microscopy, IBs can be seen as refractile particles with a diameter of 0.5–1.3 µm inside the cytoplasm of bacteria [7,8] or, for secreted proteins, in the periplasmic space [7,9–16]. Using transmission electron microscopy to observe cross-sections of cells, IBs appear, usually one per cell, as electro-dense and quite amorphous inclusions [17], although the paracrystalline structures have also been described. Following cell lysis and detergent purification, IBs have a rough surface when viewed using scanning electron microscopy [8,10]. IBs have a characteristically high density (~1.3 mg/ml^{-1}) [8,10,12,18], which allows them to be separated easily from other cellular components by using high-speed centrifugation after cell disruption [12,18]. While they are dense, IBs have a porous architecture and are highly hydrated. In fact, the release of surface proteins from IBs with limited proteolysis results in a granular architecture that suggests the existence of a complex inner structure that is formed by the clustering of smaller, protease-resistant nuclei [10]. Interestingly, several recent reports indicate that for a particular protein, the size and morphology of IBs might depend on both the culture and purification conditions, suggesting that these intracellular aggregates may have an unexpected plasticity [19]. Although in some cases, specifically at the early stages of deposition, host proteins might represent up to 50% of the IBs composition [10], in general, mature IBs contain very little host protein; on many occasions, the overexpressed protein accounts for more than 90% of the polypeptides embedded in the aggregates [20–22]. The rest of the material in the IB is likely to be ribosomal components [23], proteolytic fragments of the recombinant protein [24,25], traces of hydrophobic membrane proteins [23,26,27] or phospholipids and DNA/RNA fragments [28]. These elements might have been trapped in the IBs by way of nonspecific intermolecular interactions during the aggregation of the target protein or may have been copurified with the aggregates under low-stringency conditions [18,27]. However, some components of the cellular protein machinery, including the small heat-shock proteins IbpA and IbpB [21,29,30] and the main chaperones DnaK and GroEL [21,26,27,31], have been shown to specifically associate with IBs. DnaK is preferentially localized on the surface of IBs, suggesting a functional interaction during the solubilization of these intracellular aggregates. GroEL, which is homogeneously distributed in the cytosol, is found in low amounts inside the aggregate and is absent from the IB surface. This GroEL distribution might be functional or simply a result of co-aggregation during IB-formation. In any case, the difference in localization patterns suggests that these two proteins have different roles in defining the architectural organization of protein embedded in IBs [17].

Specificity during IB formation

The native structure of a globular protein is maintained by a delicate thermodynamic balance of hydrogen bonds, hydrophobic and electrostatic interactions. These contacts are inherently weak but are sufficient to prevent the transition into partially unfolded states that might become assembly-competent intermediates. This strategy appears to be very successful for avoiding aggregation, since few globular proteins aggregate from their stable native conformation in their natural environment. By contrast, during protein production, ongoing translation provides a continuous supply of unfolded or partially folded protein, exposing sticky hydrophobic stretches ready to establish non-native interactions with the solvent; in many cases, the result is their deposition into intracellular aggregates. Due to the lack of any noticeable pattern (sequence, structure, size or origin) between the numerous proteins able to form IBs inside prokaryotic cells, their formation has been long considered to be driven simply by the establishment of nonspecific intermolecular contacts between nascent, partially-folded species. However, it is now thought that IB formation results from protein-specific assembly [11]. Several seminal studies have provided evidence to support specificity in the *in vitro* aggregation processes of different model proteins. In this way, the study of tryptophanase refolding has shown that its aggregation depends on its own concentration and is not affected by the presence of foreign proteins; thus its aggregation is probably led by selective interactions between the enzyme polypeptide chains [32]. This specificity was also observed in the *in vitro* aggregation of a mixture of folding intermediates from P22 coat and tailspike proteins, two polypeptides that form IBs when they are overexpressed individually in bacteria but preferentially self-associate *in vitro* [33]. Importantly, our laboratory has demonstrated that purified IBs are able to capture homologous soluble proteins in a dose-dependent manner [11]. This process is conformation dependent, since proteins are recognized when they display fully or partially unfolded conformations but not once they have attained their native and stable structure. Moreover, this incorporation also seems to be sequence specific: a particular IB does not recognize unfolded, soluble heterologous polypeptide chains, suggesting that this process is directed by selective interactions between the soluble folding intermediates of a protein and its IBs [11]. Specificity has also been observed *in vivo* during the co-expression of two different proteins whose encoding genes are present on the same plasmid. In this case, two types of cytoplasmic aggregates were detected. These deposits displayed different morphologies, and electrophoretic analysis demonstrated that each of them was enriched in one type of recombinant protein [34]. In addition to conformational and sequential effects, the kinetics of protein folding and aggregation of individual aggregation-prone proteins will affect the co-aggregation of proteins. Importantly, Goloubinoff's group has shown that, *in vitro,* cross-interactions between dissimilar heat denatured proteins might also take place, suggesting that, in addition to selective contacts, cooperative interactions between low-specificity sites could also be important during protein aggregation [35].

Amyloid-like properties of IBs

The observed selectivity in IB formation is reminiscent of the behavior of amyloid aggregates. The ability to adopt an amyloid-like structure or to undergo fibrillogenesis has emerged as a common and perhaps fundamental property of polypeptide chains [36]. Amyloid formation is, in general, a nucleation-dependent process, reliant on the establishment of selective protein interactions; folding intermediates assemble in a specific manner to form discrete-structured oligomers that are expanded into prefibrillar structures and then further matured into highly ordered fibrils of a β-sheet polypeptide chain arrangement [2,37]. Similarly, IBs may also be formed by a nucleation mechanism, whereby an emerging IB acts as a nucleus to incorporate nascent proteins. The specific protein recruitment observed *in vitro*, the homogeneous composition and the low copy number inside the cell (usually one) all support this theory. IB formation and maturation also implies an enrichment of intermolecular β-structure [11,38–40]. This feature is independent of the protein's native structure and can be measured by attenuated total reflectance infrared (ATR-IR) spectroscopy [40–42]. Using this relatively simple and rapid method to probe the structural features of aggregates, Doglia and coworkers have detected a progressive increase in the cellular content of intermolecular β-sheet structure during *in vivo* formation of IBs [41]. The recurring presence of intermolecular β-structure inside IBs explains their ability to bind amyloid diagnostic dyes such as Congo Red and Thioflavin-T [11]. Rinas and coworkers have compared the cohesive forces between the polypeptides contained inside IBs with those in *in vitro* aggregates [27]. They found that both types of deposits were equally resistant to chaotropic agents and reducing conditions and more stable than the nonaggregated native protein. These results reinforce the idea that similar contacts sustain the structure of these different aggregates.

Another characteristic common to amyloids and IBs is the presence of protein regions with different sensitivity to proteolytic attack [10,43,44]. Villaverde's laboratory has studied the kinetics of trypsin digestion of purified IBs [10,43]. They observed that some fragments are immediately degraded, whereas others remain stable for a long time. Because this differential protease sensitivity could not be explained by different surface exposures of the polypeptides, the existence of different protein conformations inside IBs was proposed [10,44]. It is tempting to speculate that the resistant fraction corresponds to stable intermolecular β-sheet-rich regions in the aggregates [10,43–45]. Consistent with this, mature IBs exhibit increased resistance to proteases and higher β-sheet content than earlier forms, which could indicate an internal remodeling of the aggregates during their maturation similar to that reported for amyloids [46].

The propensity of a protein to develop *in vitro* amyloid aggregates correlates with its accumulation as IBs inside the cell, illustrating the existence of similar forces driving the formation of IBs and amyloid deposits. In *E. coli*, the expression of modified amyloidogenic proteins with reduced fibril formation propensity results in more soluble protein variants [47,48]; increasing their ability to form amyloids results in higher deposition into IBs [48–50]. In addition, *de novo* proteins designed to assemble as amyloids form IBs *in vivo* [51], whereas mutations that prevent this assembly render the proteins soluble [52]. A nice example that illustrates the relationship between *in vitro* and *in vivo* aggregation is the wild-type αA-crystallin and its G98R mutant [53].

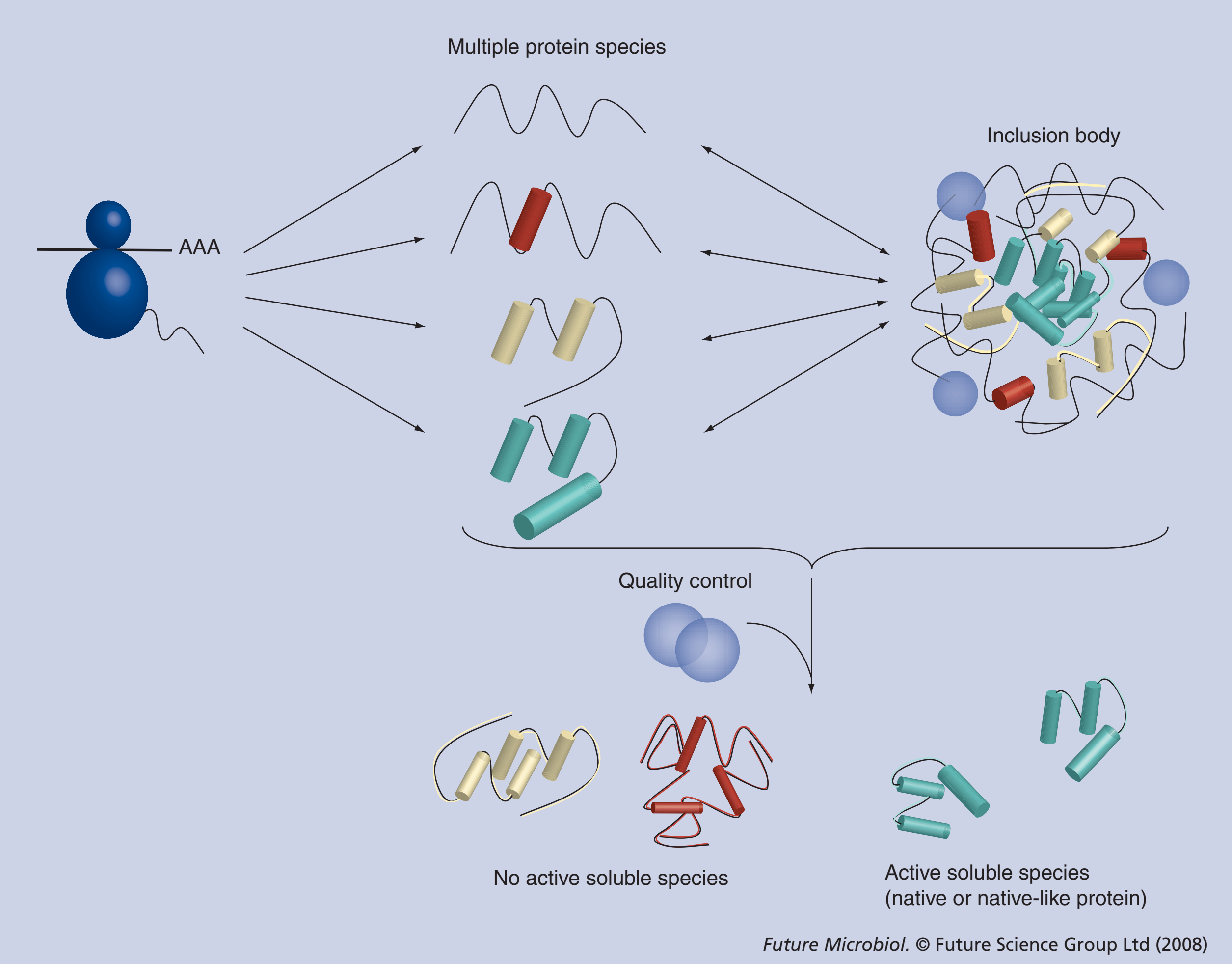
The wild-type protein remains soluble *in vitro* and is mainly expressed in the soluble fraction in bacteria; the mutated form, however, aggregates *in vitro* and is essentially localized to IBs [54]. *In vitro,* the presence of the wild-type protein in a solution containing the G98R variant reduces aggregation. In the same manner, the coexpression of these two proteins in *E. coli* inhibits IB formation and allows the majority of the expressed protein (wild-type and mutant) to accumulate in the soluble fraction [53]. Overall, these results suggest a correlation between the intermolecular interactions driving the formation of *in vitro* and *in vivo* aggregates and reinforce the existence of a specific mechanism for molecular recognition and an interaction between similar polypeptide chains inside the cell.

In the last few years, the mechanism by which amyloid aggregates cause toxicity has been the subject of intense debate. At present, the consensus view states that the prefibrillar forms are the most toxic species, whereas the final, highly-structured mature fibrils do not impair cell viability and may even play a protective role [55,56]. Accordingly, a comparison of the cytotoxicity of polyQ oligomers and their homologous eukaryotic IBs in neuronally differentiated cells reveal that cells containing IBs have longer survival periods [57]. In a similar manner, immature bacterial IBs are more toxic in both bacterial and neuronal cell lines than mature forms of the same aggregates [58,59]. Thus, the toxicity of intracellular inclusions and amyloid aggregates appear to be governed by common features related to their supramolecular structures. The ability to sequester misfolded or oligomeric forms into ordered aggregates could be a conserved generic mechanism to reduce the amount of soluble species with cytotoxic potential [55,56,58,59]. Overall, the similarities between IBs and amyloid aggregation indicate that bacterial systems can be interesting models to study the different factors modulating toxic protein deposition inside the cell [45].

IBs as reservoirs of active/native-like protein

For a long time, IBs have been considered to be inert deposits of misfolded proteins and, therefore, useless in bioprocesses. Accordingly, a large proportion of aggregation-prone products have been disregarded for commercialization. While this view is consistent with their amyloid-like nature, a number of recent independent studies have seriously challenged this assumption by demonstrating that IBs also contain native-like secondary structures. ATR-IR structural analysis of α-helical proteins, such as IL-1β, demonstrated that the IBs they form contain, in addition to the characteristic intermolecular β-structure, a significant amount of α-helix secondary conformation [60]. How these two kinds of structure are organized inside the aggregate is still not known. In any case, several recent independent investigations have detected the presence of active molecules inside IBs, which implies that aggregation of recombinant proteins in bacterial IBs does not necessarily completely inactivate them [16,61,62]. We and our coworkers have studied this feature in depth by quantifying the biological activity enclosed in the IBs formed by different enzymes and fluorescent proteins [9]. The catalytic activity and specific fluorescence exhibited by these IBs indicated that they contain a significant fraction of the total functional recombinant protein within the cell. The functionality of IBs was not homogeneous, and the aggregate core appeared to be enriched in active species [63]; still, the porosity and high hydration of IBs allowed efficient substrate diffusion to the catalytically active sites inside the aggregates [9]. The level of activity in IBs depended on the nature of the polypeptide, but, because IBs are highly pure protein microparticles, even the lowest observed functionalities are still high enough to consider the use of IBs as biocatalysts, skipping refolding procedures. The consideration of IBs as particles for industrial catalysis might lead to the rethinking of many biotechnological strategies [9]. While most protein production has aimed to increase the amount of soluble protein in the cell, little was known about the protein quality in this fraction and this protein was usually assumed to be fully functional. Recently, it has been shown that, in fact, there is a huge variety of protein conformations in the soluble fraction, including 'soluble aggregates' [64], and that the functionality of this fraction is highly variable [46,61,65]. Thus, the soluble fraction may contain some inactive or partially inactive protein forms in addition to the active, well-folded molecules [31,66]. The results reviewed in this section indicate that incorrect folding is not always paired with aggregation and correct folding cannot always be associated with solubility (Figure 1). Therefore, solubility *per se* is not the best reporter of protein quality during protein production, since the presence of active polypeptides in IBs and inactive protein forms in the soluble cell fraction result in very similar specific activities. Because the distribution of protein

Figure 1. The different structural states accessible for a recombinant protein in the cytoplasm of a bacterial cell during overexpression.



Future Microbiol. © Future Science Group Ltd (2008)

After protein synthesis, the polypeptide chain might acquire several intermediate conformational forms, from unfolded (wavy line) to the native structure (green species), which coexist in the inner cell. All these protein structures might be recruited into an inclusion body. Under the protein quality control action (blue circles), a kinetic equilibrium (double-headed arrows) between the soluble and the aggregated forms of the protein is established. As a consequence, the soluble fraction is not only composed of native protein but also soluble aggregated protein forms. This protein flux also results in enrichment of native protein in the inclusion body core. Accordingly, total soluble/insoluble protein levels are not an exact measurement of the polypeptide nativeness/inactivity during protein production.

conformers in the soluble and insoluble fractions is considerably influenced by the production conditions and/or the host cell genetic backgrounds [66], the distinction between conformational quality and solubility can be exploited in new strategies aimed to optimize quality rather than quantity during protein production.

Aggregates inside the cell: protein quality & dynamics

The conformational plasticity discussed previously is in part the result of an unbalanced, highly dynamic equilibrium between protein deposition and removal involving a continuous exchange of polypeptides between the soluble and insoluble forms of recombinant proteins [24]. In this way, if protein synthesis becomes interrupted, cytoplasmic IBs are almost totally dispersed within a few hours, since the arrest of new protein translation makes larger ratios of chaperones and foldases available to refold the precipitated protein in IBs [67]. This observation suggests that IBs may act as cellular protein reservoirs [21,46,67] from which recombinant protein can be extracted [24,68,69]. The sophisticated cell quality control system is involved in this protein flux through disaggregation, unfolding and polypeptide reactivation [46]. Chaperones, small heat shock proteins (Hsps) and proteases are the main components of this control system, which is activated under stress situations, such as heat-shock or protein overexpression [70–74]. According to their mode of action, the different elements of the cell quality control machinery can be classified as unfolders, folders and holders [74]. Folder chaperones, mainly the DnaK/Hsp70 family and GroEL [75], have the ability to refold misfolded or aggregated proteins [74]. Sometimes folder chaperones require a previously unfolded substrate to act [76,77], in which case they work together with unfolder chaperones, principally ATPase-associated chaperones [74]. Finally, the holder activity protects polypeptides against aggregation [74]. This function is executed by small Hsps, which are able to bind aggregates [30] and folding intermediates and isolate them from the crowded environment [75], allowing other elements of the quality control system to act [70,78–80]. These activities all have a remodeling effect on the structure and composition of IBs. In this sense, it has been suggested that the removal of misfolded protein from the surface of IBs [81,82] results in a progressive enrichment of the native-like and active protein in the IB core [63]. Villaverde and coworkers have measured and compared the distribution of active protein in the soluble and insoluble fractions in bacterial knockouts of several chaperones [31]. In general, total or partial removal of chaperones causes a decrease in the amount of soluble protein and an increase in the β-sheet compactness inside IBs. Surprisingly, total activity of both soluble and insoluble fractions increased, while the cellular proteolytic activity decreased, indicating that chaperones can modulate the digestion of partially folded intermediates, even if they maintain a certain active conformation. Therefore, the quality control system improves protein solubility, while at the same time sacrificing functionality.

The results of two genomic initiatives to rationally understand the bacterial reaction to the development of intracellular insoluble aggregates have been recently reported [6,83]. Genes expressed during the synthesis of aggregation-prone proteins were compared with those expressed during the production of almost completely soluble polypeptides. Transcriptional profiles for multiple examples in the soluble and insoluble classes were used to identify patterns of gene expression that correlate with protein solubility. In response to translational misfolding, expression of the heat shock sigma factor ρ^{32} target genes are elevated. The same group of genes was induced by the expression of insoluble protein under different growth conditions, which likely reflects a generalized cellular response to protein insolubility. The response is functional, in that nearly every component of the protein folding machinery is found in this set of genes and overexpression of chaperones was consistently detected in both studies, suggesting that the bacterial host responds to protein aggregation by increasing its global folding capacity [6,83]. Also, the expression of ribosome-associated genes was altered [83]. Modulation of the translational activity might provide an effective measure against aggregation by holding the emerging protein in the relatively protected environment of the translating ribosome until sufficient chaperone molecules can be recruited. By identifying a minimal set of genes responding to insoluble protein accumulation, it becomes possible to globally or selectively control gene expression as an alternative and potentially general strategy for improving the solubility of recombinant proteins.

Sequential & structural determinants of protein aggregation in bacteria

The development of robust strategies to control protein aggregation requires a deep understanding of both extrinsic and intrinsic determinants dictating protein deposition. Apart from the cellular mechanisms described previously, specific sequential and conformational characteristics of proteins modulate their propensities to aggregate. Accordingly, we and others have shown that for a given aggregation-prone protein, there are certain sequence segments (hot spots) that specifically assist in its deposition [48,84–87]. These regions are usually protected in the native protein structure, but become exposed during protein production when there is a high population of a variety of partially folded protein conformers [64]. At least some of these folding intermediates would present uncovered hot spots able to self-assemble and nucleate protein deposition [84,87]. These stretches are preferential targets for tackling protein aggregation. To determine whether the sequence in these regions modulates *in vivo* protein aggregation reactions, our group has extensively mutated a hot spot of the amyloid-β peptide and characterized the aggregation propensities of these different variants inside *E. coli.* The measured tendencies for aggregation *in vivo* correlate with variations in the intrinsic properties of the polypeptide that are promoted by the different mutations [48]. Specifically, as previously shown *in vitro* [88], the hydrophobicity, propensity to form β-sheet secondary structure and the charge of the protein appear to modulate intracellular aggregation. Sequential changes in this region not only control the proportion of soluble and aggregated recombinant protein but also the degree of functionality of the corresponding IBs [89]. Remarkably, the level of activity inside the IBs significantly correlates with the predicted aggregation rates for the different variants. This result suggests that there is a kinetic competition between protein folding and aggregation during recombinant expression [89]. In other words, the conformational quality of IBs is determined by the time needed to achieve a correct folding prior to the aggregation event. For a given polypeptide with a constant folding rate, the faster the protein aggregates, the lower its activity and *vice versa*, so that active molecules in IBs are those whose aggregation was slow enough to permit prior polypeptide folding. It is known that the folding routes of proteins displaying slow folding rates usually imply the transient accumulation of metastable folding intermediates [90]. These protein species would display uncovered sticky regions that might favor their deposition and inactivation if the aggregation rate is high enough, suggesting that slow-folding polypeptides would have higher aggregation propensity than fast-folding ones. In related work, Balaji and coworkers studied the correlation between protein stability and aggregation propensity, comparing a set of polypeptide properties between different groups of proteins (soluble proteins, IB-forming proteins and amyloidogenic proteins) [91]. They found that the set of soluble proteins displayed on average a lower contact order. The contact order parameter, defined as the normalized average sequence separation between interacting residues in the folded state, is a measure of local versus long-range interactions in the native-state structure [92]. This parameter is small for proteins that are stabilized mainly by local interactions and is large for proteins whose residues interact frequently with partners that are far away in the protein sequence. Hence, the proteins in the soluble set, with a smaller contact order, are predicted to fold faster than IB-forming proteins. Interestingly enough, the same study demonstrates that soluble proteins also possess fewer exposed hydrophobic residues, higher helix propensity and less-charged polar residues; all three of these properties, according to our experimental data, effectively reduce aggregation propensity. In good agreement, Chiti and coworkers have shown that, whereas destabilized mutants of the N-terminal domain of the *E. coli* protein HypF invariably aggregate after expression, the aggregation of destabilized variants can be prevented by increasing the net charge of the protein [93]. Thus, an understanding of the intrinsic factors that govern protein deposition can now be exploited to design new, more soluble, protein variants with either accelerated folding rates or decreased aggregation propensity.

How to reduce protein aggregation

In biotechnological protein production, there are two main strategies used to increase protein yields: improve *in vivo* native folding during protein synthesis or optimize protein purification from IBs [94]. They can be applied synergistically to maximize protein production, but unfortunately, as described previously, both depend on the particular characteristics of the protein of interest, leading to a time- and resource-expensive search for each individual optimal expression and purification condition. As a result, an overwhelmingly vast amount of research has been devoted in the last few years to investigating generic ways to increase protein yields.

One of the key steps for structural genomics and proteomics is high-throughput expression of target proteins, of which usually only a small fraction ends up being soluble. The most common strategy to modulate protein aggregation is to shift culture growth conditions by modifying temperature, growth medium richness, the inducer, its concentration or the induction time. Usually, several of these parameters need to be adjusted. The complex interplay between them requires exploring a large number of combinations before optimal conditions are found, and yet there is still little information available on how these factors influence IB characteristics. Our group has recently shown that the temperature of cultivation influences, in a almost linear manner, both the amount of active protein inside the IBs and the conformational stability of the aggregates [95]. Data suggest that folding into a native conformation stabilized by intrachain contacts and aggregation into a non-native structure stabilized by interchain interactions are competing processes. Higher cultivation temperatures favor the formation of intermolecular interactions, rather than the native intramolecular ones, leading to an increase in IB stability but a significant decrease in their functionality. By contrast, IBs formed at low temperatures are highly functional but less stable. This result suggests that it would be, in principle, possible to obtain active protein from these IBs using simple and economic downstream strategies, such us mild IB disaggregation, without the need for aggressive and usually inefficient unfolding–refolding steps.

Another successful strategy in protein production makes use of molecular chaperones by co-expressing these proteins together with the target. Bukau and coworkers have recently tested the effect of overproducing the entire network of major cytosolic chaperones [96]. They used a two-step strategy: first, the chaperones are coproduced with the recombinant protein to enhance folding of the nascent polypeptide; second, biosynthesis is interrupted by inhibiting the continuous generation of novel aggregation-prone proteins and favoring the chaperone-assisted folding. With this approach, most of the recombinant proteins tested (70% of 64 different heterologous proteins) displayed decreased accumulation in IBs [96]. However, one must keep in mind that in some instances the chaperones actually exacerbate rather than ameliorate the formation of IBs. It has been suggested that small, organic molecules acting as chemical chaperones could be used as effective, alternative, agents in preventing protein aggregation. They are compounds that preferentially stabilize the native protein conformation during biosynthesis or refolding. In this way, the amino acid arginine slows down the refolding reaction and at the same time reduces the interactions between folding intermediates, thus decreasing aggregation and favoring the formation of the correct conformation [97]. Osmoprotectants have also been shown to have a protective effect against aggregation [98–100]. In particular, the effect of proline on protein stability and aggregation has been recently studied using a fluorescent reporter designed to monitor *in vivo* and *in vitro* protein deposition [100]. Proline turns out to have a solphobic impact on the protein backbone that destabilizes forms with a high amount of solvent-accessible surface area (like partially folded states) and favors species with less exposed surface (like the native structure). Thus, adding this amino acid at the early stages of production stabilizes the native state and disfavors the population of folded intermediates, precluding the establishment of intermolecular contacts that might lead to aggregation [100]. Unfortunately, most of the aforementioned strategies to improve protein solubility do not scale easily for high-throughput expression screening. Therefore, designing approaches to monitor the influence of the aforementioned factors, genes and compounds on *in vivo* protein deposition in a straightforward manner is of great interest. In this context, de Marco's group has developed a smart approach by using a fusion of the IbpAB gene promoter to the β-galatosidase enzyme engineered by Lesley and coworkers [83]. This fusion acts as a reporter of the conformational state of the recombinant product because this promoter is upregulated in response to protein aggregation [101]. This approach allowed testing of the influence of growth temperature, induction conditions, overexpression of chaperones, presence of osmolytes in the culture or the use of different expression vectors, simply by measuring the intracellular β-galatosidase levels.

Although we described many extrinsic factors that can be tuned to modulate protein deposition under a given set of experimental conditions, aggregation is ultimately an individual trait of proteins that is determined by their specific amino acid sequence [48,85,86,102]. Using AGGRESCAN, an in-house algorithm designed to forecast the overall protein aggregation propensity of proteins from their sequence, we found that proteins shown to be soluble under overexpression conditions in *E. coli* displayed, on average, lower predicted aggregation tendencies than their IB-forming counterparts [103]. This kind of approach might allow for the theoretical prediction of solubility from sequence that would enable scientists to rationally identify natural and designed proteins with high solubility upon overexpression, rather than resort to the trial and error procedures that are presently used. Accordingly, two recent support vector machine-based algorithms developed by the groups of Balaji and Frishman reported accuracies higher than 70% in deciphering whether a particular protein sequence would be soluble or form IBs when overexpressed [104,105]. They also identified a subset of features that have the strongest impact on protein aggregation. Among them, aliphatic index and charge promote solubility, whereas hydrophobic residues promote aggregation, a result concordant with the data obtained experimentally. An obvious limitation of these sequence-based approaches is that they do not consider key contributors in governing the solubility status of the proteins, such as the structures and stabilities of the folding intermediates and that of the native protein. Nevertheless, it is obvious that these and related programs provide a starting point for the rational engineering of protein solubility and are likely to become important tools in large-scale structural genomics initiatives.

Future perspective

The information presented here demonstrates that the investigation of the factors modulating aggregation into IBs is undergoing a rapid and very productive phase of growth. In the next few years, it is very likely that we will witness the construction of a highly synergic environment in which the integration of structural, physiological, genome-wide and *in silico* studies, using systems biology approaches, might allow modeling and understanding of the global process of protein deposition in bacteria. Such a model would describe

Executive summary

Introduction

- Protein aggregation is linked to several human diseases and also constitutes an important concern in biotechnology.
- Bacteria are widely used as factories for the production of recombinant polypeptides. Nevertheless, they display an intrinsic propensity to accumulate heterologous products in insoluble inclusion bodies (IBs).
- New relevant data on IBs would allow for the delineation of an integrated model of intracellular protein aggregation.

What is an IB? Anatomy & composition

- IBs appear as refractile particles in the bacterial cytoplasm or periplasmic space.
- They are dense, insoluble but porous and highly hydrated protein aggregates.
- IBs are highly homogeneous in composition but molecular chaperones usually localize specifically to them.

Specificity during IB formation

- IB formation seems to depend on the specific self-assembly of polypeptides through selective intermolecular contacts.

Amyloid-like properties of IBs

- IB formation and maturation implies enrichment in intermolecular β-sheet secondary structure.
- The propensity of a protein to develop *in vitro* amyloid aggregates correlates with its accumulation as IBs inside the cell.
- Bacterial IBs are toxic for both prokaryotic and neuronal cells.

IBs as reservoirs of active/native-like protein

- Aggregation of recombinant proteins can occur, as bacterial IBs do not necessarily inactivate them.
- IBs might be used as functional microparticles for biocatalysis.
- *In vivo* protein conformational quality and solubility are not coincidental traits.

Aggregates inside the cell: protein quality & dynamics

- IBs are not inert protein aggregates. They result from an unbalanced, highly dynamic equilibrium between protein deposition and removal.
- The cells quality control system sharply modulates this protein flux through disaggregation, unfolding and polypeptide reactivation processes.

Sequential & structural determinants of protein aggregation in bacteria

- Certain sequence segments (hot spots) specifically assist protein deposition. These stretches constitute preferential targets for anti-aggregational strategies.
- There is a kinetic competition between protein folding and aggregation during recombinant protein expression.
- Folding speed and conformational stability modulate protein aggregation.

How to reduce protein aggregation

- The use of chaperones, osmolytes or fusion proteins, together with the control of culture conditions, constitute the most commonly employed strategies to reduce IB formation during protein production.
- Computational approaches are useful tools to predict protein aggregation and thus, for the rational engineering of protein solubility.

Future perspective

- The integration of structural, physiological, genome-wide and *in silico* data should allow an accurate modeling and precise understanding of the global dynamic control of protein aggregation in bacteria. This can have a profound effect on the efficiency of biotechnological processes and probably improve our present knowledge on the determinants of protein deposition in the deleterious human conformational diseases.

essential features of biological interactions between cell components and with the physicochemical environment and, more importantly, predict how these interactions will evolve in structure and function. This advancement of the field should have a profound effect on the efficiency of biotechnology operations by allowing rational design of recombinant protein expression and biocatalysts. In addition, one can speculate that a detailed scenario describing the dynamic control of protein aggregation in intracellular backgrounds can be exploited to develop new drugs and therapies to fight conformational diseases, which in many cases are also the outcome of undesired protein misfolding and aggregation in human tissues.

Acknowledgements

Apologies to all those researchers whose excellent work could not be cited.

Financial & competing interests disclosure

This work was supported by grant BIO2007–68046 from the Ministerio de Ciencia y Tecnología, Spain, and by grant 2005SGR–00037 from AGAUR, Catalonia. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Bibliography

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Smith A: Protein misfolding. *Nature* 426, 883 (2003).
2. Dobson CM: Protein folding and misfolding. *Nature* 426, 884–890 (2003).
3. Rochet JC, Lansbury PT Jr: Amyloid fibrillogenesis: themes and variations. *Curr. Opin. Struct. Biol.* 10, 60–68 (2000).
4. Cohen FE, Kelly JW: Therapeutic approaches to protein misfolding diseases. *Nature* 426, 905–909 (2003).
5. Walsh DM, Selkoe DJ: A β-oligomers: a decade of discovery. *J. Neurochem.* 101, 1172–1184 (2007).
6. Smith HE: The transcriptional response of *Escherichia coli* to recombinant protein insolubility. *J. Struct. Funct. Genomics* 8, 27–35 (2007).

•• Identification of a pattern of gene expression that correlates strongly with protein solubility. This work provides a starting point for the rational design of growth parameters and host strains with improved protein solubility characteristics.

7. Carrio MM, Corchero JL, Villaverde A: Dynamics of *in vivo* protein aggregation: building inclusion bodies in recombinant bacteria. *FEMS Microbiol. Lett.* 169, 9–15 (1998).
8. Bowden GA, Paredes AM, Georgiou G: Structure and morphology of protein inclusion bodies in *Escherichia coli. Biotechnology (NY)* 9, 725–730 (1991).
9. Garcia-Fruitos E, Gonzalez-Montalban N, Morell M *et al.*: Aggregation as bacterial inclusion bodies does not imply inactivation of enzymes and fluorescent proteins. *Microb. Cell Fact.* 4, 27 (2005).

• Evidence of the presence of functional and properly folded polypeptides inside inclusion bodies.

10. Carrio MM, Cubarsi R, Villaverde A: Fine architecture of bacterial inclusion bodies. *FEBS Lett.* 471, 7–11 (2000).
11. Carrio M, Gonzalez-Montalban N, Vera A, Villaverde A, Ventura S: Amyloid-like properties of bacterial inclusion bodies. *J. Mol. Biol.* 347, 1025–1037 (2005).

•• Description of the specificity of protein assembly during inclusion body formation and characterization of the structural properties of these aggregates.

12. Taylor G, Hoare M, Gray DR, Marston FAO: Size and density of protein inclusion bodies. *Biotechnology (NY)* 4, 553–557 (1986).
13. Georgiou G, Telford JN, Shuler ML, Wilson DB: Localization of inclusion bodies in *Escherichia coli* overproducing β-lactamase or alkaline phosphatase. *Appl. Environ. Microbiol.* 52, 1157–1161 (1986).
14. Bowden GA, Georgiou G: Folding and aggregation of β-lactamase in the periplasmic space of *Escherichia coli. J. Biol. Chem.* 265, 16760–16766 (1990).
15. Miot M, Betton JM: Protein quality control in the bacterial periplasm. *Microb. Cell Fact.* 3, 4 (2004).
16. Arie JP, Miot M, Sassoon N, Betton JM: Formation of active inclusion bodies in the periplasm of *Escherichia coli. Mol. Microbiol.* 62, 427–437 (2006).
17. Carrio MM, Villaverde A: Localization of chaperones DnaK and GroEL in bacterial inclusion bodies. *J. Bacteriol.* 187, 3599–3601 (2005).
18. Georgiou G, Valax P: Isolating inclusion bodies from bacteria. *Methods Enzymol.* 309, 48–58 (1999).
19. Peternel S, Jevsevar S, Bele M, Gaberc-Porekar V, Menart V: New properties of inclusion bodies with implications for biotechnology. *Biotechnol. Appl. Biochem.* 49(Pt 4), 239–246 (2007).

• Aquisition of soluble protein from inclusion bodies without refolding steps.

20. Villaverde A, Carrio MM: Protein aggregation in recombinant bacteria: biological role of inclusion bodies. *Biotechnol. Lett.* 25, 1385–1395 (2003).
21. Carrio MM, Villaverde A: Construction and deconstruction of bacterial inclusion bodies. *J. Biotechnol.* 96, 3–12 (2002).

• Demonstration of reversibility during inclusion body formation.

22. Clark ED: Protein refolding for industrial processes. *Curr. Opin. Biotechnol.* 12, 202–207 (2001).
23. Rinas U, Bailey JE: Protein compositional analysis of inclusion bodies produced in recombinant *Escherichia coli. Appl. Microbiol. Biotechnol.* 37, 609–614 (1992).
24. Carrio MM, Corchero JL, Villaverde A: Proteolytic digestion of bacterial inclusion body proteins during dynamic transition between soluble and insoluble forms. *Biochim. Biophys. Acta* 1434, 170–176 (1999).
25. Corchero JL, Viaplana E, Benito A, Villaverde A: The position of the heterologous domain can influence the solubility and proteolysis of β-galactosidase fusion proteins in *E. coli. J. Biotechnol.* 48, 191–200 (1996).
26. Jurgen B, Lin HY, Riemschneider S *et al.*: Monitoring of genes that respond to overproduction of an insoluble recombinant protein in *Escherichia coli* glucose-limited fed-batch fermentations. *Biotechnol. Bioeng.* 70, 217–224 (2000).

27. Rinas U, Hoffmann F, Betiku E, Estape D, Marten S: Inclusion body anatomy and functioning of chaperone-mediated *in vivo* inclusion body disassembly during high-level recombinant protein production in *Escherichia coli*. *J. Biotechnol.* 127, 244–257 (2007).

• Description of inclusion body composition and plasticity.

28. Valax P, Georgiou G: Molecular characterization of β-lactamase inclusion bodies produced in *Escherichia coli*. 1. Composition. *Biotechnol. Prog.* 9, 539–547 (1993).

29. Hoffmann F, Rinas U: Kinetics of heat-shock response and inclusion body formation during temperature-induced production of basic fibroblast growth factor in high-cell-density cultures of recombinant *Escherichia coli*. *Biotechnol. Prog.* 16, 1000–1007 (2000).

30. Allen SP, Polazzi JO, Gierse JK, Easton AM: Two novel heat shock genes encoding proteins produced in response to heterologous protein expression in *Escherichia coli*. *J. Bacteriol.* 174, 6938–6947 (1992).

31. Garcia-Fruitos E, Martinez-Alonso M, Gonzalez-Montalban N, Valli M, Mattanovich D, Villaverde A: Divergent genetic control of protein solubility and conformational quality in *Escherichia coli*. *J. Mol. Biol.* 374, 195–205 (2007).

• Demonstration that the cell quality control system promotes protein solubility but not necessarily conformational quality.

32. London J, Skrzynia C, Goldberg ME: Renaturation of *Escherichia coli* tryptophanase after exposure to 8 M urea. Evidence for the existence of nucleation centers. *Eur. J. Biochem.* 47, 409–415 (1974).

33. Speed MA, Wang DI, King J: Specific aggregation of partially folded polypeptide chains: the molecular basis of inclusion body composition. *Nat. Biotechnol.* 14, 1283–1287 (1996).

• Seminal work demonstrating specificity during protein aggregation processes.

34. Hart RA, Rinas U, Bailey JE: Protein composition of *Vitreoscilla* hemoglobin inclusion bodies produced in *Escherichia coli*. *J. Biol. Chem.* 265, 12728–12733 (1990).

35. Ben-Zvi AP, Goloubinoff P: Review: mechanisms of disaggregation and refolding of stable protein aggregates by molecular chaperones. *J. Struct. Biol.* 135, 84–93 (2001).

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

37. Harper JD, Lansbury PT Jr: Models of amyloid seeding in Alzheimer's disease and scrapie: mechanistic truths and physiological consequences of the time-dependent solubility of amyloid proteins. *Annu. Rev. Biochem.* 66, 385–407 (1997).

38. Fink AL: Protein aggregation: folding aggregates, inclusion bodies and amyloid. *Fold. Des.* 3, R9–R23 (1998).

•• Pioneering review anticipating several important features of protein deposition.

39. Przybycien TM, Dunn JP, Valax P, Georgiou G: Secondary structure characterization of β-lactamase inclusion bodies. *Protein Eng.* 7, 131–136 (1994).

40. Ami D, Natalello A, Gatti-Lafranconi P, Lotti M, Doglia SM: Kinetics of inclusion body formation studied in intact cells by FT-IR spectroscopy. *FEBS Lett.* 579, 3433–3436 (2005).

41. Ami D, Natalello A, Taylor G, Tonon G, Maria Doglia S: Structural analysis of protein inclusion bodies by Fourier transform infrared microspectroscopy. *Biochim. Biophys. Acta* 1764, 793–799 (2006).

• Application of infrared spectroscopy to the characterization of the conformational properties of inclusion bodies.

42. Ami D, Bonecchi L, Cali S, Orsini G, Tonon G, Doglia SM: FT-IR study of heterologous protein expression in recombinant *Escherichia coli* strains. *Biochim. Biophys. Acta* 1624, 6–10 (2003).

43. Cubarsi R, Carrio MM, Villaverde A: *In situ* proteolytic digestion of inclusion body polypeptides occurs as a cascade process. *Biochem. Biophys. Res. Commun.* 282, 436–441 (2001).

44. Cubarsi R, Carrio MM, Villaverde A: A mathematical approach to molecular organization and proteolytic disintegration of bacterial inclusion bodies. *Math. Med. Biol.* 22, 209–226 (2005).

45. Ventura S, Villaverde A: Protein quality in bacterial inclusion bodies. *Trends Biotechnol.* 24, 179–185 (2006).

46. Gonzalez-Montalban N, Garcia-Fruitos E, Ventura S, Aris A, Villaverde A: The chaperone DnaK controls the fractioning of functional protein between soluble and insoluble cell fractions in inclusion body-forming cells. *Microb. Cell Fact.* 5, 26 (2006).

47. Wigley WC, Stidham RD, Smith NM, Hunt JF, Thomas PJ: Protein solubility and folding monitored *in vivo* by structural complementation of a genetic marker protein. *Nat. Biotechnol.* 19, 131–136 (2001).

48. de Groot NS, Aviles FX, Vendrell J, Ventura S: Mutagenesis of the central hydrophobic cluster in Aβ42 Alzheimer's peptide. Side-chain properties correlate with aggregation propensities. *FEBS J.* 273, 658–668 (2006).

49. Sirangelo I, Malmo C, Casillo M, Mezzogiorno A, Papa M, Irace G: Tryptophanyl substitutions in apomyoglobin determine protein aggregation and amyloid-like fibril formation at physiological pH. *J. Biol. Chem.* 277, 45887–45891 (2002).

50. Hammarstrom P, Sekijima Y, White JT *et al.*: D18G transthyretin is monomeric, aggregation prone, and not detectable in plasma and cerebrospinal fluid: a prescription for central nervous system amyloidosis? *Biochemistry* 42, 6656–6663 (2003).

51. West MW, Wang W, Patterson J, Mancias JD, Beasley JR, Hecht MH: *De novo* amyloid proteins from designed combinatorial libraries. *Proc. Natl Acad. Sci. USA* 96, 11211–11216 (1999).

52. Wang W, Hecht MH: Rationally designed mutations convert *de novo* amyloid-like fibrils into monomeric β-sheet proteins. *Proc. Natl Acad. Sci. USA* 99, 2760–2765 (2002).

53. Singh D, Raman B, Ramakrishna T, Rao Ch M: Mixed oligomer formation between human αA-crystallin and its cataract-causing G98R mutant: structural, stability and functional differences. *J. Mol. Biol.* 373, 1293–1304 (2007).

• Important study demonstrating that similar forces govern *in vitro* and *in vivo* specificity during protein aggregation.

54. Singh D, Raman B, Ramakrishna T, Rao ChM: The cataract-causing mutation G98R in human αA-crystallin leads to folding defects and loss of chaperone activity. *Mol. Vis.* 12, 1372–1379 (2006).

55. Bucciantini M, Giannoni E, Chiti F *et al.*: Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature* 416, 507–511 (2002).

56. Marianayagam NJ, Sunde M, Matthews JM: The power of two: protein dimerization in biology. *Trends Biochem. Sci.* 29, 618–625 (2004).

57. Takahashi T, Kikuchi S, Katada S, Nagai Y, Nishizawa M, Onodera O: Soluble polyglutamine oligomers formed prior to inclusion body formation are cytotoxic. *Hum. Mol. Genet.* 17, 345–356 (2008).

58. Gonzalez-Montalban N, Villaverde A, Aris A: Amyloid-linked cellular toxicity triggered by bacterial inclusion bodies. *Biochem Biophys. Res. Commun.* 355, 637–642 (2007).

- Characterization of the toxicity of bacterial inclusion bodies.

59. Gonzalez-Montalban N, Carrio MM, Cuatrecasas S, Aris A, Villaverde A: Bacterial inclusion bodies are cytotoxic *in vivo* in absence of functional chaperones DnaK or GroEL. *J. Biotechnol.* 118, 406–412 (2005).

60. Oberg K, Chrunyk BA, Wetzel R, Fink AL: Nativelike secondary structure in interleukin-1 β inclusion bodies by attenuated total reflectance FTIR. *Biochemistry* 33, 2628–2634 (1994).

61. Garcia-Fruitos E, Carrio MM, Aris A, Villaverde A: Folding of a misfolding-prone β-galactosidase in absence of DnaK. *Biotechnol. Bioeng.* 90, 869–875 (2005).

62. Tokatlidis K, Dhurjati P, Millet J, Beguin P, Aubert JP: High activity of inclusion bodies formed in *Escherichia coli* overproducing *Clostridium thermocellum* endoglucanase D. *FEBS Lett.* 282, 205–208 (1991).

63. Garcia-Fruitos E, Aris A, Villaverde A: Localization of functional polypeptides in bacterial inclusion bodies. *Appl. Environ. Microbiol.* 73, 289–294 (2007).

64. Schrodel A, de Marco A: Characterization of the aggregates formed during recombinant protein expression in bacteria. *BMC Biochem.* 6, 10 (2005).

•• Demonstrates the complexity of the aggregation pattern of a recombinant protein expressed in bacteria and characterizes the different aggregate subclasses.

65. Vera A, Gonzalez-Montalban N, Aris A, Villaverde A: The conformational quality of insoluble recombinant proteins is enhanced at low growth temperatures. *Biotechnol. Bioeng.* 96, 1101–1106 (2007).

66. Gonzalez-Montalban N, Garcia-Fruitos E, Villaverde A: Recombinant protein solubility: does more mean better? *Nat. Biotechnol.* 25, 718–720 (2007).

67. Carrio MM, Villaverde A: Protein aggregation as bacterial inclusion bodies is reversible. *FEBS Lett.* 489, 29–33 (2001).

68. Vera A, Aris A, Carrio M, Gonzalez-Montalban N, Villaverde A: Lon and ClpP proteases participate in the physiological disintegration of bacterial inclusion bodies. *J. Biotechnol.* 119, 163–171 (2005).

69. Corchero JL, Cubarsi R, Enfors S, Villaverde A: Limited *in vivo* proteolysis of aggregated proteins. *Biochem. Biophys. Res. Commun.* 237, 325–330 (1997).

70. Mogk A, Deuerling E, Vorderwulbecke S, Vierling E, Bukau B: Small heat shock proteins, ClpB and the DnaK system form a functional triade in reversing protein aggregation. *Mol. Microbiol.* 50, 585–595 (2003).

71. Mogk A, Bukau B: Molecular chaperones: structure of a protein disaggregase. *Curr. Biol.* 14, R78–R80 (2004).

- Structure and disaggregation mechanism of the chaperone ClpB.

72. Mogk A, Schlieker C, Strub C, Rist W, Weibezahn J, Bukau B: Roles of individual domains and conserved motifs of the AAA$^+$ chaperone ClpB in oligomerization, ATP hydrolysis, and chaperone activity. *J. Biol. Chem.* 278, 17615–17624 (2003).

73. Weibezahn J, Bukau B, Mogk A: Unscrambling an egg: protein disaggregation by AAA$^+$ proteins. *Microb. Cell Fact.* 3, 1 (2004).

74. Dougan DA, Mogk A, Bukau B: Protein folding and degradation in bacteria: to degrade or not to degrade? That is the question. *Cell. Mol. Life Sci.* 59, 1607–1616 (2002).

75. Liberek K, Lewandowska A, Zietkiewicz S: Chaperones in control of protein disaggregation. *EMBO J.* 27, 328–335 (2008).

- Recent review on the molecular mechanisms by which chaperones liberate and refold polypeptides trapped in protein aggregates.

76. Doyle SM, Hoskins JR, Wickner S: Collaboration between the ClpB AAA$^+$ remodeling protein and the DnaK chaperone system. *Proc. Natl Acad. Sci. USA* 104, 11138–11144 (2007).

- Elegant study on the role played by different chaperones in protein disaggregation.

77. Lewandowska A, Matuszewska M, Liberek K: Conformational properties of aggregated polypeptides determine ClpB-dependence in the disaggregation process. *J. Mol. Biol.* 371, 800–811 (2007).

78. Mogk A, Schlieker C, Friedrich KL, Schonfeld HJ, Vierling E, Bukau B: Refolding of substrates bound to small Hsps relies on a disaggregation reaction mediated most efficiently by ClpB/DnaK. *J. Biol. Chem.* 278, 31033–31042 (2003).

79. Lee GJ, Vierling E: A small heat shock protein cooperates with heat shock protein 70 systems to reactivate a heat-denatured protein. *Plant Physiol.* 122, 189–198 (2000).

80. Matuszewska M, Kuczynska-Wisnik D, Laskowska E, Liberek K: The small heat shock protein IbpA of *Escherichia coli* cooperates with IbpB in stabilization of thermally aggregated proteins in a disaggregation competent state. *J. Biol. Chem.* 280, 12292–12298 (2005).

81. Schlieker C, Tews I, Bukau B, Mogk A: Solubilization of aggregated proteins by ClpB/DnaK relies on the continuous extraction of unfolded polypeptides. *FEBS Lett.* 578, 351–356 (2004).

82. Schlieker C, Weibezahn J, Patzelt H *et al.*: Substrate recognition by the AAA$^+$ chaperone ClpB. *Nat. Struct. Mol. Biol.* 11, 607–615 (2004).

83. Lesley SA, Graziano J, Cho CY, Knuth MW, Klock HE: Gene expression response to misfolded protein as a screen for soluble recombinant protein. *Protein Eng.* 15, 153–160 (2002).

84. Ivanova MI, Sawaya MR, Gingery M, Attinger A, Eisenberg D: An amyloid-forming segment of β2-microglobulin suggests a molecular model for the fibril. *Proc. Natl Acad. Sci. USA* 101, 10584–10589 (2004).

85. Ventura S: Sequence determinants of protein aggregation: tools to increase protein solubility. *Microb. Cell Fact.* 4, 11 (2005).

- Review describing the modulation of protein aggregation by sequential traits.

86. Ventura S, Lacroix E, Serrano L: Insights into the origin of the tendency of the PI3-SH3 domain to form amyloid fibrils. *J. Mol. Biol.* 322, 1147–1158 (2002).

87. Ventura S, Zurdo J, Narayanan S *et al.*: 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 (2004).

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

89. de Groot NS, Ventura S: Protein activity in bacterial inclusion bodies correlates with predicted aggregation rates. *J. Biotechnol.* 125, 110–113 (2006).

- Presents one of the few available evidences for *in vivo* kinetic competition between folding and aggregation.

90. Nolting B, Schalike W, Hampel P *et al.*: Structural determinants of the rate of protein folding. *J. Theor. Biol.* 223, 299–307 (2003).
91. Idicula-Thomas S, Balaji PV: Correlation between the structural stability and aggregation propensity of proteins. *In Silico Biol.* 7, 225–237 (2007).
    - Theoretical work suggesting a correlation between the conformational stability of polypeptides and their tendency to aggregate, which can have predictive utilities.
92. Plaxco KW, Simons KT, Baker D: Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* 277, 985–994 (1998).
93. Calloni G, Zoffoli S, Stefani M, Dobson CM, Chiti F: Investigating the effects of mutations on protein aggregation in the cell. *J. Biol. Chem.* 280, 10607–10613 (2005).
94. Qoronfleh MW, Hesterberg LK, Seefeldt MB: Confronting high-throughput protein refolding using high pressure and solution screens. *Protein Expr. Purif.* 55, 209–224 (2007).
95. de Groot NS, Ventura S: Effect of temperature on protein quality in bacterial inclusion bodies. *FEBS Lett.* 580, 6471–6476 (2006).
96. de Marco A, Deuerling E, Mogk A, Tomoyasu T, Bukau B: Chaperone-based procedure to increase yields of soluble recombinant proteins produced in *E. coli*. *BMC Biotechnol.* 7, 32 (2007).
    - Exhaustive study on the effect of chaperones on the solubility of proteins during recombinant expression.
97. Liu YD, Li JJ, Wang FW, Chen J, Li P, Su ZG: A newly proposed mechanism for arginine-assisted protein refolding: not inhibiting soluble oligomers although promoting a correct structure. *Protein Expr. Purif.* 51, 235–242 (2007).
98. Diamant S, Rosenthal D, Azem A, Eliahu N, Ben-Zvi AP, Goloubinoff P: Dicarboxylic amino acids and glycine–betaine regulate chaperone-mediated protein-disaggregation under stress. *Mol. Microbiol.* 49, 401–410 (2003).
99. de Marco A, Vigh L, Diamant S, Goloubinoff P: Native folding of aggregation-prone recombinant proteins in *Escherichia coli* by osmolytes, plasmid- or benzyl alcohol-overexpressed molecular chaperones. *Cell Stress Chaperones* 10, 329–339 (2005).
100. Ignatova Z, Gierasch LM: Inhibition of protein aggregation *in vitro* and *in vivo* by a natural osmoprotectant. *Proc. Natl Acad. Sci. USA* 103, 13357–13361 (2006).
101. Schultz T, Martinez L, de Marco A: The evaluation of the factors that cause aggregation during recombinant expression in *E. coli* is simplified by the employment of an aggregation-sensitive reporter. *Microb. Cell Fact.* 5, 28 (2006).
102. Sanchez de Groot N, Pallares I, Aviles FX, Vendrell J, Ventura S: Prediction of 'hot spots' of aggregation in disease-linked polypeptides. *BMC Struct. Biol.* 5, 18 (2005).
103. Conchillo-Sole O, de Groot NS, Aviles FX, Vendrell J, Daura X, Ventura S: AGGRESCAN: a server for the prediction and evaluation of 'hot spots' of aggregation in polypeptides. *BMC Bioinformatics* 8, 65 (2007).
104. Idicula-Thomas S, Kulkarni AJ, Kulkarni BD, Jayaraman VK, Balaji PV: A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in *Escherichia coli*. *Bioinformatics* 22, 278–284 (2006).
105. Smialowski P, Martin-Galiano AJ, Mikolajka A, Girschick T, Holak TA, Frishman D: Protein solubility: sequence-based prediction and experimental verification. *Bioinformatics* 23, 2536–2542 (2007).

Affiliations

- *Natalia S de Groot*
  Departament de Bioquímica i Biologia Molecular & Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, E-08193 Bellaterra, Spain
- *Alba Espargaró*
  Departament de Bioquímica i Biologia Molecular & Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, E-08193 Bellaterra, Spain
- *Montserrat Morell*
  Departament de Bioquímica i Biologia Molecular & Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, E-08193 Bellaterra, Spain
- *Salvador Ventura*
  Departament de Bioquímica i Biologia Molecular & Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, E-08193 Bellaterra, Spain
  Tel.: +34 935 868 147;
  Fax: +34 935 811 264;
  salvador.ventura@uab.es

Review

Amyloids in bacterial inclusion bodies

Natalia S. de Groot*, Raimon Sabate* and Salvador Ventura

Departament de Bioquímica i Biologia Molecular and Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain

Protein misfolding and aggregation into amyloid structures are associated with dozens of human diseases. Recent studies have provided compelling evidence for the existence of highly ordered, amyloid-like conformations in the insoluble inclusion bodies produced during heterologous protein expression in bacteria. Thus, amyloid aggregation seems to be an omnipresent process in both eukaryotic and prokaryotic organisms. Amyloid formation inside cell factories raises important safety concerns with regard to the toxicity and infectivity of recombinant proteins. Yet such findings also suggest that prokaryotic cells could be useful systems for studying how and why proteins aggregate *in vivo*, and they could also provide a biologically relevant background for screening therapeutic approaches to pathologic protein deposition.

Protein aggregation

Protein aggregation stems from the self-association of identical polypeptides to form insoluble, higher-order assemblies that ultimately precipitate. This self-assembly process has attracted the attention of many scientists over the past few years, in part because a connection exists between the formation of insoluble protein deposits in tissues and the development of more than 40 different human diseases, many of which are debilitating and fatal [1]. Additionally, protein aggregation in bacteria forms a major bottleneck in the protein production pipeline and has narrowed the spectrum of protein-based drugs that are available in the biotechnology market [2]. Disease-related protein aggregation is usually characterized by the formation of highly ordered, long, straight and unbranched amyloid fibrils that share a common cross-β-sheet structure (see Glossary) [1,3]. By contrast, protein deposition in bacteria occurs in the form of inclusion bodies (IBs), which are conventionally regarded as amorphous aggregates [4]. As these supramolecular structures have different macroscopic morphologies, the molecular mechanisms underlying protein deposition in eukaryotic and prokaryotic organisms have long been thought to be unrelated.

Although significant effort has been devoted to understanding the fine structure and molecular mechanisms underlying amyloid fibril formation, relatively little attention has been paid to bacterial protein aggregates. Several lines of evidence suggest, however, that on a microscopic level, IBs might share several common features with the highly structured amyloid fibrils that are associated with human disorders [5,6]. Nevertheless, obtaining a detailed structural characterization of these prokaryotic intracellular deposits has proven to be extremely challenging, and the degree to which IBs resemble amyloids remained essentially unclear. Recently, however, independent studies have provided solid evidence for a common mechanism underlying fibril formation in conformational disorders and the aggregation of proteins inside bacterial cells [7,8]. In this review, we will discuss how these data have challenged our view of IBs, as well as the biotechnological and biological implications of these discoveries.

The common view of IB formation, structure and function

Globular proteins rarely aggregate in their biological environments. The maintenance of their native conformation depends on a delicate balance of non-covalent forces that are intrinsically weak but sufficient to provide proteins with structural uniqueness. By contrast, protein production in cell factories occurs with a high translation rate and thereby provides the cell with a continuous supply of unfolded polypeptides. In this case, both the hydrophobic side-chains and the backbone hydrogen bond donors/acceptors are exposed to the solvent and ready to interact, either intramolecularly to funnel the protein towards the functional structure or intermolecularly to form aggregated species. Under these conditions, first-order folding and second-order aggregation reactions readily compete inside the cell [9]. In many cases of recombinant expression, where high intracellular protein concentrations are achieved, aggregation dominates over folding and insoluble protein deposits are formed. The stability of the aggregated form is often higher than that of the native structure [10]; therefore, this form acts as a thermodynamic and kinetic trap from which it is difficult for embedded polypeptides to escape [9]. Because IBs form during the expression of a large number of proteins that are seemingly unrelated in sequence, structure, size or origin, they have been long thought to grow as a result of the formation of nonspecific contacts between polypeptide chains as they emerge from ribosomes. In support of this theory, all IBs share a common amorphous appearance, regardless of the target protein [11]. They are very dense refractile particles that can be found in both the cytoplasmic and periplasmic space of bacteria. They can be nearly 1 μm in diameter and do not seem to have any ordered inner structure. Because the contacts leading to the formation of IBs were thought to be non-native and nonspecific, these aggregates have been assumed to contain mostly misfolded polypeptides that have no regular secondary structure and are consequently devoid of any

Corresponding author: Ventura, S. (salvador.ventura@uab.es)

[*] These authors contributed equally..

0968-0004/$ – see front matter © 2009 Elsevier Ltd. All rights reserved. doi:10.1016/j.tibs.2009.03.009 Available online xxxxxx

Glossary

[PSI$^+$]: the misfolded form of Sup35, which is an important protein factor for translation termination during protein synthesis. It is believed that [PSI$^+$] causes suppression of nonsense mutations by sequestering functional Sup35p in non-functional aggregates, thereby allowing stop codon readthrough.

Aggresome: a proteinaceous inclusion body formed around the microtubule-organizing centre in eukaryotic cells when the degradation machinery is impaired or overwhelmed. Its formation involves a retrograde, microtubule-based transport and is largely believed to be a protective response, sequestering potentially cytotoxic aggregates.

Amide I region: region of the infrared spectrum corresponding to the vibration of the amide bond and comprising a frequency of 1600–1700 cm^{-1}. Its line shape is sensitive to the type and amount of secondary structures and is not strongly influenced by side chains.

Cell factories: refers to bacterial cell systems used for large-scale production of different valuable biological products, usually of proteic nature.

Cellular quality-control machinery: the group of cellular components responsible for maintenance of protein homeostasis by catalysing refolding and/or the immediate destruction of misfolded or impaired proteins generated in cells.

Chaperone: a protein that assists the non-covalent folding/unfolding and the assembly/disassembly of other proteins or proteic macromolecular structures.

Circular dichroism (CD) spectroscopy: measures differences in the absorption of left-handed polarized light versus right-handed polarized light that arise due to structural asymmetry.

Cross-β-sheet structure: the characteristic amyloid fibril structural pattern. In this quaternary structure the β-sheets are parallel to the fibril axis and the β-strands within a sheet are perpendicular to the fibril axis.

First-order reaction: in this context, a first-order reaction indicates that, for monomeric proteins, the mole fraction of protein folded under particular conditions is independent of the concentration.

Fluorescence resonance energy transfer (FRET): non-radioactive energy transfer phenomena between two fluorophores. A previously excited donor fluorophore might transfer photons to an acceptor fluorophore at distances of a few nm, depending on the spectral overlap and proper dipole alignment of the two fluorophores.

FRET efficiency: the fraction of photons absorbed by the donor fluorophore that is transferred to the acceptor fluorophore. This value depends on the proximity of the donor and acceptor molecules.

Hot spot: in this context, a short protein region with particular physicochemical properties (e.g. high hydrophobicity and/or β-sheet propensity) that, if exposed to solvent, might initiate the aggregation process.

Hydrogen–deuterium exchange (H/D exchange): a chemical reaction in which a covalently bonded hydrogen atom is replaced by a deuterium atom, or vice versa. Hydrogen exchange measurements can be used to sense changes in protein structure on a specific timescale or to differentiate protein regions that display different exposure to the solvent.

Infrared (IR) spectroscopy: because chemical bonds absorb infrared energy at specific frequencies (or wavelengths), the basic structure of compounds can be determined by the spectral locations of their infrared absorptions.

Intrinsically unstructured protein: proteins that lack a stable and well-defined tertiary structure but, despite this, remain biologically functional.

Oligomer: small, non-covalently bound, metastable multimer formed at the early stages of the aggregation pathway. It is usually considered as a spherical or elliptical assembly that precedes the formation of protofibrillar structures.

Prefibrillar assemblies: oligomers and protofibrils.

Prion protein: an infectious and transmissible amyloid or amyloid-like assembly capable of self-replicating its conformation *in vivo and in vitro*.

Protofibril: Non-spherical, 'rod-like' or 'worm-like' filamentous structures that are devoid of a regular periodic substructure and that represent intermediates in the formation of highly ordered amyloid fibrils.

Pulse–chase: experimental procedure that tracks the fates of proteins through a cell, from the protein's synthesis to its final cellular destination. In these experiments, cells are grown in radioactive medium for a brief period (the pulse) and then transferred to non-radioactive medium for a longer period (the chase).

Refractile: refers to a particle within the cell that scatters (refracts) light.

Second-order reaction: in this context, a second-order reaction indicates that the mole fraction of an aggregated protein under particular conditions is dependent on the initial protein concentration.

Seed: a preformed and stable aggregate that provides a scaffold for rapid amyloid elongation.

functional activity. Active protein can be recovered from these deposits through successive unfolding and refolding steps. However, the recovery is usually low and the procedure requires adaptation for each target protein. In general, IBs were considered to be inert cellular 'dust balls' that were of little interest to either the basic or applied sciences [2].

Challenging the classical model: amyloid structure in IBs

The presence of a cross-β structure in amyloid fibrils was broadly accepted long before high-resolution structural data could confirm that polypeptides in these supramolecular assemblies adopt an extended β-sheet conformation, with the β-strands stacked perpendicularly to the long axis of the fibrils [1,3]. Lower resolution approaches, including infrared (IR) spectroscopy, circular dichroism (CD) spectroscopy and low-resolution X-ray diffraction, were first used to show that fibrils formed by structurally and sequentially unrelated proteins share an enhancement in β-sheet content [1]. Likewise, the secondary structure of bacterial IBs has been analysed with IR, and a significant increase in β-sheet structure, relative to the functional conformation and independent of the native structure, is observed in all proteins that have been assayed to date. The IR spectra of IBs in the amide I region is dominated by a main signal of around 1620 cm^{-1} (Figure 1). This band indicates the formation of a new, extended, intermolecular β-sheet conformation with polypeptide backbones that are tightly packed through short hydrogen bonds; this structure is very similar to the cross-β structure present in amyloids [7,12]. In many IBs, however, there is also a detectable presence of disordered conformations and, in some cases, native-like secondary structure [13–15]. X-ray diffraction data of IBs strongly support the presence of amyloid-like contacts in conformationally unrelated proteins. They all display reflections at 4.7 Å, which is consistent with the spacing between strands in a β-sheet, and at ~10 Å, which is interpreted as the distance between adjacent β-sheets (Figure 1) [8]. Again, these two reflections are characteristic of amyloid fibrils [1,3]. Nevertheless, their circular profiles indicate that, in contrast to fibrils, these structures are not strongly aligned [8]. The CD spectrum of IBs in the far-UV region also indicates the dominance of a β-sheet secondary structure, which usually coexists with apparently unstructured polypeptide conformations (Figure 1) [8]. Finally, the cross-β-sheet motif is thought to be the main structural element responsible for the specific binding of the dyes thioflavin-T (Th-T) and congo red to amyloid fibrils. These two dyes also bind IBs (Figure 1); in fact, in some cases they bind with a higher affinity to IBs than to pathogenic amyloid fibrils [5,8].

Conventionally, for a protein aggregate to be considered an amyloid, it must meet three requirements: it forms fibrils that are visible in electron microscopy (EM) or atomic force microscopy (AFM); it binds Th-T and congo red; and it has a high β-sheet content [3]. Although IBs satisfy the last two conditions, they still appear amorphous on the macroscopic scale. Amyloids do not necessarily incorporate their entire polypeptide length into the highly packed β-sheet structure that constitutes the core of the fibrils [16]; the rest of the protein remains disordered or even in a globular and active conformation, as seen in yeast prions [17]. The comparative analysis of the secondary structure content of fibrils and IBs has shown that non

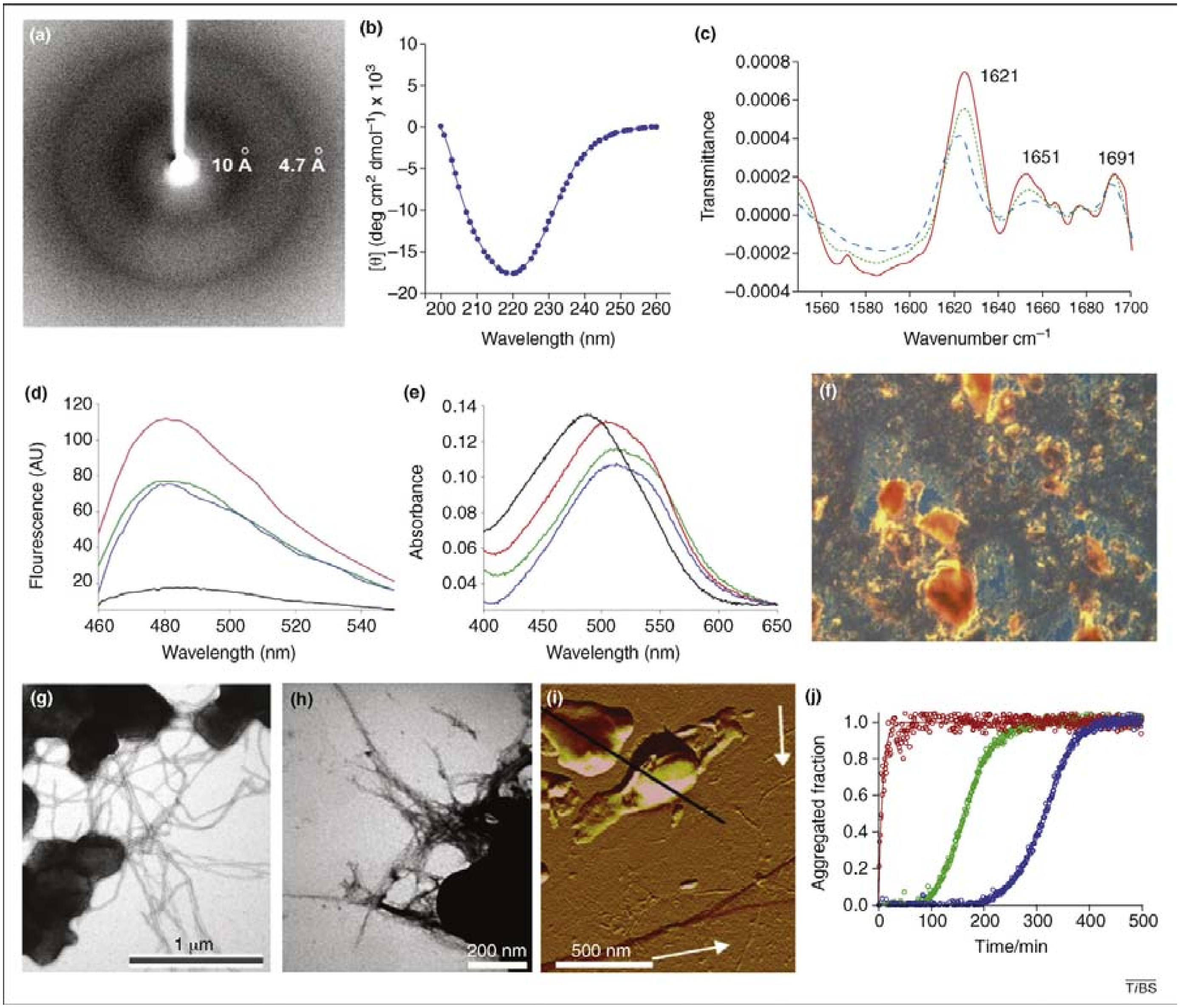


Figure 1. Amyloid properties of bacterial IBs. The biophysical and structural characterization of the IBs formed by unrelated proteins demonstrates that these insoluble deposits contain amyloid-like structures. β-Sheet secondary structures in IBs: **(a)** X-ray diffraction of early secreted antigen 6-kDa protein (ESAT-6) IBs, showing the two typical reflections at 4.7 Å and ~10 Å consistent with a cross-β structure; **(b)** far-UV CD spectra of ESAT-6 IBs, displaying the characteristic minimum at around 217 nm indicative of a β-sheet conformation; **(c)** Fourier transform infrared spectroscopy (FTIR) spectrum of the IBs formed by two fusions of β-galactosidase and the capsid protein of foot-and-mouth disease virus (VP1LAC, red, and LACVP1, blue) and the tailspike protein (TSP, green), with two bands at 1621 cm^{-1} and 1691 cm^{-1} characteristic of a intermolecular β-sheet structure; the band at 1651 cm^{-1} indicates the presence of disordered conformations.Binding of IBs to amyloid-specific dyes: **(d)** increase in the fluorescence emission of Th-T in the presence of different IBs (VP1LAC in green, LACVP1 in red and TSP in blue) compared with that promoted by soluble conformations (VP1LAC in black); **(e)** shift in the absorption spectra of congo red in the presence of different IBs (VP1LAC in green, LACVP1 in red and TSP in blue) compared with that promoted by soluble conformations (VP1LAC in black); **(f)** congo red birefringence of the IBs formed by myelin oligodendrocyte glycoprotein (MOG) under cross-polarized light. Amyloid-like fibrils in IBs imaged by transmission electron microscopy (TEM) and atomic force microscopy (AFM): **(g)** TEM image of bone morphogenetic protein-2 IBs after *in vitro* incubation at 37 °C for 12 h; **(h)** TEM image of Aβ IBs after 30 min of PK proteolytic action; **(i)** AFM image of IBs (Aβ–GFP) digested for 30 min with PK; fibril-like structures settling on the graphite surface are indicated with white arrows. Seeding-dependent formation of amyloid fibrils: **(j)** the formation of Aβ fibrils is accelerated by the addition of preformed fibrils (red line) and Aβ IBs (green line) in comparison to the spontaneous aggregation of monomeric Aβ (blue line). (a), (b), (f) and (g) adapted, with permission, from Ref. [8]; (c), (d) and (e) adapted, with permission, from Ref. [5]; and (h), (i) and (j) adapted, with permission, from Ref. [7].

cross-β regions are much more abundant in bacterial aggregates. This composition, together with their high protein density (1.3 mg/ml) [18], could prevent the observation of fibrillar structure in IBs. Fortunately, it is possible to discriminate between fibrillar and non-fibrillar regions in protein aggregates. Proteinase K (PK) has been extensively used to map the core of amyloid fibrils because it is highly active against globular or disordered conformations but displays low activity against densely packed cross-β regions [19]. Interestingly, PK treatment of the apparently amorphous bacterial IBs formed by the Alzheimer's disease (AD) amyloid β (Aβ) peptide promotes the appearance of elongated fibrillar structures with heights ranging from 2 to 5 nm, which is consistent with the dimensions and morphology of amyloid protofilaments and fibrils as observed both by EM and AFM (Figure 1). Accordingly, PK-digested material binds Th-T and congo red with higher affinity than intact IBs [7].

The presence of a structured core in a protein aggregate can also be mapped by measuring quenched

Table 1. Common structural and functional characteristics of amyloid aggregates and bacterial IBs

Amyloid	Refs	Inclusion body	Refs
Typically comprises a single primary protein	[1]	>90% is constituted by the recombinant protein	[24]
Nucleation-polymerization fibrillization process	[20]	Can act as seeds for protein aggregation	[5]
Heterologous co-aggregation is rare	[21]	Different co-expressed proteins do not co-aggregate	[7,53]
Sequence-dependent aggregation	[1]	Sequence-specific aggregation	[5]
Amino acid changes in aggregation-prone regions affect aggregation kinetics	[54]	Single amino acid changes in 'hot spots' strongly affect IB deposition	[8,30]
Chaperone activity regulates aggregation	[55]	IBs interact with the cellular quality-control machinery	[11]
Aggregation stems from partially unfolded intermediates	[1,38]	IB formation correlates with the population of partially unfolded states	[5,37]
Stabilization of the native state decreases aggregation	[38]	Intrinsic stability inversely correlates with IB formation	[13,37]]
Detection of cross-β-sheet intermolecular organization by FTIR and CD	[1]	Detection of β-sheet structure by FTIR and CD	[7,8,12]
Cross-β-sheet X-ray diffraction pattern	[56]	Cross-β-sheet X-ray diffraction pattern	[8]
β-sheet protected core observed by NMR	[57]	Preferentially protected regions with β-sheet structure detected by NMR	[8]
Binding of congo red and Th-T	[1,3]	Binding of congo red and Th-T	[5]
Fibrillar structures observed by AFM and TEM	[1,3]	Presence of inner fibrillar structure coexisting with amorphous material	[7]
Protease-resistant regions that match the fibrillar core	[19]	Regions with preferential proteolytic resistance	[7]
Different fibrillar structures display different stability	[1]	IBs from different proteins have different stability	[34]
Cross-β and globular structures might coexist	[17]	IBs could contain globular/active conformations	[13,14]
Prions are toxic for mammalian cells	[58]	Recombinant prions are toxic for bacteria	[51]
Initial soluble oligomers are SDS-stable	[45]	Presence of SDS-stable oligomers after induction	[7]
Cytotoxic prefibrillar assemblies	[45]	IBs are toxic against mammalian cells	[49]
Mature fibrils are less cytotoxic	[47]	Toxicity inversely correlates with cross-β presence	[49]
Might affect cell division and aging	[59]	Influence bacterial division and aging	[50]

hydrogen–deuterium (H/D) exchange using solution nuclear magnetic resonance (NMR) because amide protons involved in strong hydrogen bonds in the β-core become protected from the solvent. Using this approach, Riek and coworkers [8] have analysed the IBs formed by three proteins belonging to different structural classes. In all three proteins, they detected the presence of cross-β-sheet structures surrounded by disordered regions, in agreement with the PK data. In contrast to Aβ, which is an intrinsically unstructured protein, this study investigated globular polypeptides that are not associated with any disease, indicating that the formation of amyloid-like structures inside IBs might be a general phenomenon [8]. Therefore, it seems that the establishment of an inter-backbone, hydrogen-bonded network that stabilizes related fibrillar structures enriched in the β-sheet conformation is a common force driving protein aggregation *in vivo* (Table 1). Thus, amorphous aggregates, as defined to date, simply might not exist. Although the presence of amyloid stretches in bacterial IBs might seem surprising, it can explain most, if not all, of the properties of these intracellular aggregates.

Sequential determinants of IB formation

The formation of fibrils by an amyloidogenic protein is accelerated by the presence of preformed fibrils or seeds [20]. This seeding behaviour is thought to promote the fast development of AD after its clinical detection. Amyloid seeding is usually sequence-specific: aggregation is nucleated by homologous fibrils but not by fibrils from closely related sequences [21]. Prions, however, are a remarkable exception; here, cross-seeding allows trespassing across the species barrier [22]. In conformational diseases, amyloids typically contain a single primary protein rather than a mixture of polypeptides that were nonspecifically recruited to the aggregate. Interestingly, IBs are also highly enriched in the recombinant target, which can constitute up to 90% of the total mass of the aggregate [23]. In addition, it is very common to find only one IB per cell, suggesting that a reduced number of aggregation nuclei exist at early stages [24]. One of these nuclei would then grow by the continuous incorporation of the monomeric target polypeptide. We have provided support for this hypothesis by demonstrating that, *in vitro*, purified IBs recognize and incorporate at their surface homologous, but not heterologous, polypeptides in a dose-dependent manner [5].

It is now widely accepted that specific continuous protein segments nucleate the aggregation reaction and participate in the formation of the β-core of the mature fibrils [25]. Aggregation-prone regions have been identified in most of the polypeptides that underlie neurodegenerative and systemic amyloidogenic disorders [1]. Accordingly, different computational approaches have been developed to accurately predict those stretches of sequence in the fibrils of distinct pathogenic proteins [26–28]. The presence of an amyloid-core in IBs suggests that similar sequences might also be responsible for the contacts that lead to the selective incorporation of homologous polypeptide chains during IBs formation. In support of this idea, the central hydrophobic cluster (CHC), including residues 17–21, is the most protected region in Aβ IBs [7]. This particular region has a key role in Aβ fibril formation and is located in the core of the fibrils [29]. The similarity between the β-core of fibrils and IBs explains why the polypeptides embedded in Aβ IBs can specifically recognize soluble Aβ monomers and accelerate their fibrillization [7] (Figure 1). The ability of IBs to seed amyloid formation is perhaps the most compelling evidence to support the amyloid nature of IBs. Additionally, the cross-β segments in IBs that are formed by non-disease-related globular proteins are computationally predicted to possess a high amyloidogenic propensity [8]. Those regions, usually known as 'hot spots', typically comprise up to ten residues, have a high intrinsic aggregation propensity and are compatible with a densely packed β-sheet conformation [8,26–28]. When these regions, which correspond to the detected cross-β regions in IBs, are synthesized chemically as short peptides, they readily form typical amyloid fibrils [8]. The presence of a single 'hot spot' seems to be sufficient to mediate the aggregation of the entire polypeptide into IBs [8].

Because IB formation relies on specific contacts between short, selective and predictable regions, the introduction of aggregation-disrupting amino acid substitutions in these sequences invariably increases the solubility of the target proteins [8,30]. This result explains the recurrent observation that the same changes in different proteins can differentially impact solubility [31]; it is now clear that the impact of the substitution depends crucially on the protein region where it occurs. Importantly, these discoveries provide an opportunity to predict and fine-tune protein solubility during recombinant expression in bacteria by selectively modifying the primary sequence [2]. Interestingly, organisms tend to favour high expression of less aggregation-prone proteins [32].

This new view of IBs, which we call the 'IB-stretch hypothesis', suggests that the strength of the interactions that hold the polypeptides inside the aggregates would be unique for each protein. In fact, we have shown that the IBs formed by different proteins have specific thermodynamic and kinetic stability features, explaining why some IBs are easily disaggregated, whereas others require high concentrations of denaturants. Because the separation of individual polypeptide chains from the aggregate is a rate-limiting step for the action of molecular chaperones on aggregated species [33], IBs of different proteins must impose dissimilar challenges to the cellular quality-control machinery [34].

The 'IB-stretch hypothesis' posits that IBs might also contain globular and functional domains, provided that the crucial residues for the active protein conformation are not engaged in the β-core of the aggregate. This idea is consistent with the observation that the globular domains of wild-type and engineered yeast prions remain functional in amyloid fibrils when attached downstream of the prion-determining sequence [17]. Accordingly, an increasing number of proteins are reported to be at least partially active inside bacterial aggregates [13,14]. Hence, IBs might not require a refolding step to be directly used for biotechnology purposes.

Conformational determinants of IB formation

In intrinsically unstructured proteins such as Aβ, 'hot spots' are exposed to the solvent and are ready to establish intermolecular contacts that might ultimately lead to their aggregation [35]. By contrast, globular proteins that form IBs contain aggregation-prone sequences that map to regular secondary structure elements in the native conformation, thus preventing their direct exposure to the solvent [8]. Accordingly, IBs do not recruit properly folded homologous polypeptides and thus aggregation in bacteria requires globular proteins to be at least partially unfolded [5]. Because unfolded and partially folded intermediates accumulate after protein synthesis in the ribosome, it has been thought that the bulk of protein aggregation occurs during the time between translation and the acquisition of the native structure (Figure 2). The *in vivo* aggregation of a polypeptide can be monitored by fusing it to a functional reporter, for example green fluorescent protein (GFP) [7,13,14] (Box 1). We fused GFP to Aβ variants with different aggregation propensities and showed that there is indeed a kinetic competition between folding and aggregation. Fast aggregating sequences generated poorly fluorescent IBs because the GFP failed to fold before its aggregation. By contrast, the IBs of slow aggregating variants were highly fluorescent [13]. Our data indicate, however, that aggregation does not occur immediately

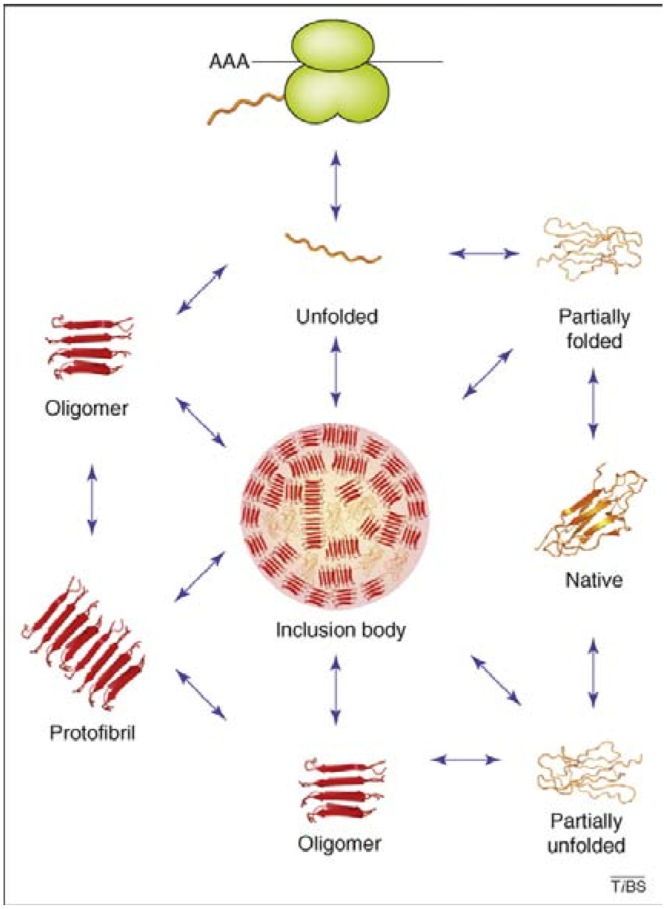


Figure 2. Protein conformations that lead to IB formation. Nascent polypeptide chains in the ribosome (green) are, in principle, devoid of regular structure and fold into the soluble, monomeric, native conformation through selective intermolecular contacts. This process often involves the population of one or more partially folded intermediates. However, during protein production, both unfolded and partially folded conformers often establish anomalous, but specific, intermolecular interactions. These species might be directly sequestered by homologous sequences in preformed IBs (orange), which act as nuclei for aggregation. They can also self-assemble into oligomeric and protofibrillar structures that might themselves act as seeds for the formation of new IBs or be incorporated in pre-existing aggregates. Properly folded counterparts in the cytoplasm are not completely saved from aggregation because fluctuations in the structure might promote local unfolding and self-assembly through previously hidden, aggregation-prone regions. Even if this aggregation stems from poorly populated non-native states, it might deplete the native conformation in a time-dependent manner. The populations and interconversions of the various states depend on their relative thermodynamic and kinetic stabilities in the cell. For all involved species, self-assembly promotes the enrichment in β-sheet structure.

Box 1. Aggregation specificity and kinetics during IB formation

The recurrent presence of cross-β regions within specific sequence stretches in IBs suggests that bacterial intracellular aggregation is a selective process. The analysis of co-aggregation between homologous and heterologous self-aggregating proteins is an elegant way to test the specificity of the aggregation process. In their seminal work, Hart and coworkers [53] show that the simultaneous co-expression of two different heterologous proteins in the same bacterial cell results in the formation of cytoplasmic aggregates differing in their relative content of recombinant proteins. More recently, we labelled two aggregation-prone proteins with different fluorescent tags and co-expressed them in *E. coli* [7]. As a negative control, we co-expressed two differentially labelled versions of one protein. In both cases, typically, a single IB accumulated at one of the cells' poles. This approach allows the visualization of each type of polypeptide in the aggregate (Figure I), as well as the ability to infer their proximity using fluorescence resonance energy transfer (FRET). In the control IBs, the two fluorescent signals co-localize and display a high FRET efficiency, indicating self-aggregation of the two tagged proteins. By contrast, the associated fluorescence signals do not significantly co-localize in the IBs formed by deposition of two different polypeptides. One protein was embedded in the inner core of the aggregate, whereas the other one decorated the outer face (Figure I). Importantly, even in areas where the two fluorescence signals apparently overlapped, the FRET efficiency is low, indicating that donor and acceptor molecules, and accordingly their fused self-aggregating polypeptides, are not close in space and therefore are unlikely to interact at the molecular level.

In mixed IBs, the relative position of the two different polypeptides, in or out, is determined by their relative aggregation propensities (Figure I); the faster aggregating protein excluded the other from the core of the IBs. This kinetic segregation confirms that each polypeptide establishes preferential interactions with the homologous sequences during aggregation and therefore that the two proteins do not effectively co-aggregate. These behaviours are highly similar to those of proteins that aggregate in the cytoplasm and nucleus of mammalian cells into aggresomes and nuclear inclusions [74,75], respectively. Such findings suggest the existence of evolutionarily conserved adaptations in eukaryote and prokaryote cells that control the deposition of misfolded proteins.

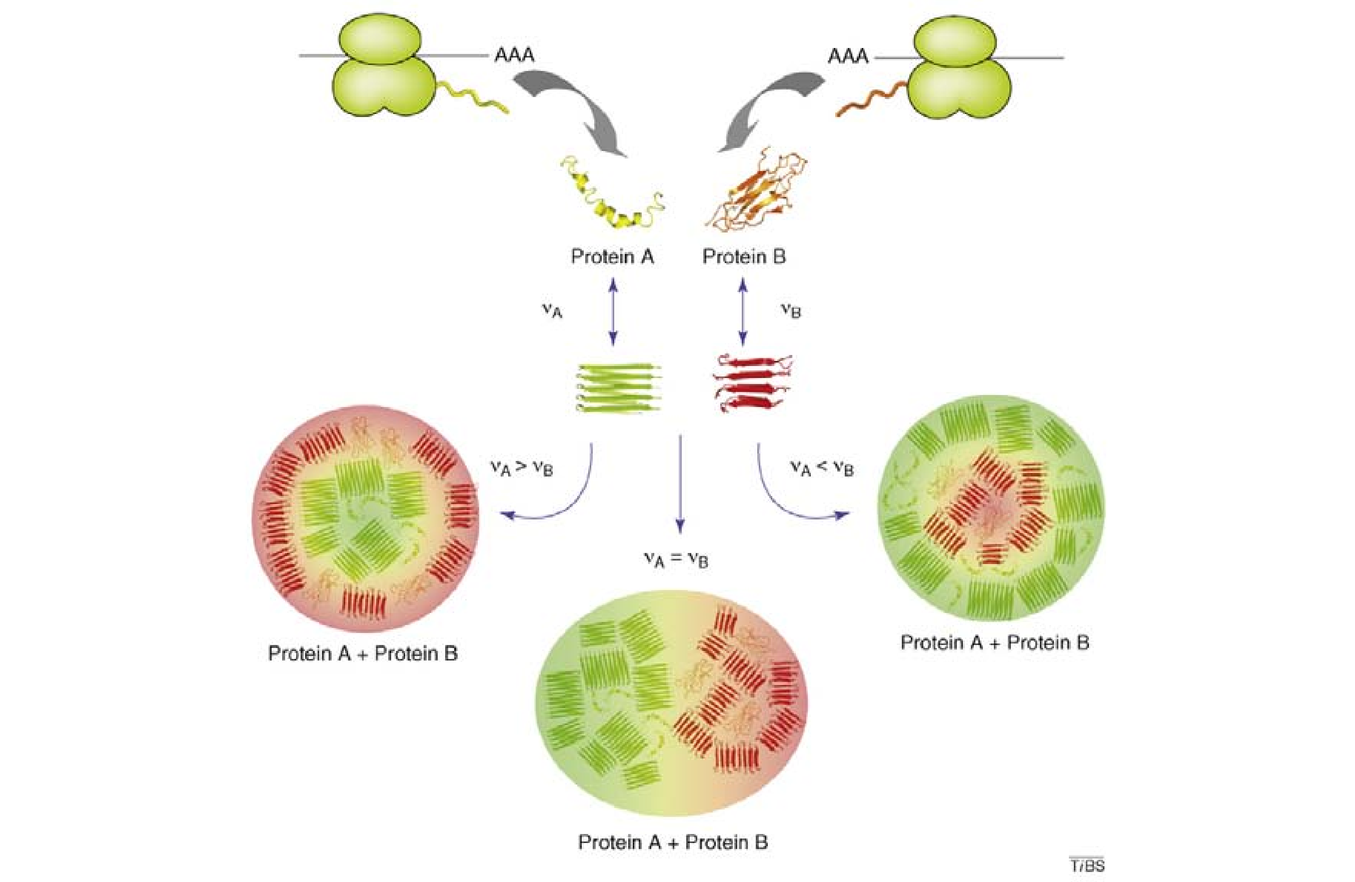


Figure I. Specificity and kinetic segregation of polypeptide chains during their aggregation into IBs. The scheme illustrates IB formation during the co-expression of two different proteins in the same cell. The two proteins self-assemble through selective interactions between identical polypeptide chains. Their relative aggregation rates (ν) determine their relative position in the aggregate; the protein with the highest aggregation propensity occupies the core of the aggregate. When the two aggregation rates are similar, sequence specificity is expected to result in protein compartmentalization in the mature aggregate.

after synthesis. Pulse–chase experiments showed that some globular proteins remain susceptible to aggregation at time periods well beyond the time required to fold in the bacterial cytoplasm [36]. Moreover, many proteins form IBs even though they are able to fold in the millisecond time-scale without populating intermediates [37]. It seems possible, therefore, that globular proteins in bacteria can also aggregate under conditions in which they are initially folded. One possibility is that fluctuations in the native state result in local unfolding and that the transient exposure of aggregation-prone regions allows them to interact with preformed IBs or homologous nascent polypeptide chains (Figure 2).

Recently, this mechanism was proposed for the formation of pathogenic amyloid fibrils by globular proteins under physiological conditions [38]. An important implica-

Table 2. Amyloidogenic proteins involved in human diseases that accumulate as insoluble IBs when expressed recombinantly in bacteria

Polypeptide[a]	Disease	Native structure	Refs reporting IB formation
α-Crystallin (G98R mutant)	Cataracts	All β	[60]
Amyloid-β peptide	Alzheimer's disease	Natively unfolded	[7]
Amylin (IAPP)	Type II diabetes	Natively unfolded	[61]
Atrial natriuretic factor	Atrial amyloidosis	Natively unfolded	[62]
N-terminal fragments of apolipoprotein AI	ApoAI amyloidosis	Natively unfolded	[63]
β2-Microglobulin	Haemodialysis-related amyloidosis	All β, Ig-like	[64]
Calcitonin	Medullary carcinoma of the thyroid	Natively unfolded	[65]
Cystatin C (L68Q variant)	Hereditary cystatin C amyloid angiopathy	α+β, cystatin-like	[66]
Human prion protein or fragments	Spongiform encephalopathies	Natively unfolded	[67]
Huntingtin with polyQ expansion	Huntington's disease	Natively unfolded	[68]
Immunoglobulin light chains or fragments	Amyloid light chain amyloidosis	All β, Ig-like	[69]
Insulin	Injection-localized amyloidosis	All α, insulin-like	[70]
Keratins	Cutaneous lichen amyloidosis	Unknown	[71]
Superoxide dismutase1	Amyotrophic lateral sclerosis	All β, Ig-like	[72]
Mutants of transthyretin (D18G, V30M, L55P)	Familial amyloidotic polyneuropathy	All β, prealbumin-like	[43]
Lysozyme	Lysozyme amyloidosis	α+β, lysozyme fold	[73]

[a]The amyloid properties of these polypeptides as well as their conformation and the associated disorders are described in detail in [1].

tion of this hypothesis is that the intrinsic stability of a protein might modulate its *in vivo* aggregation; thus, destabilization of the native state would increase conformational fluctuations, whereas over-stabilization would decrease the presence of transiently exposed 'hot spots' [37,38]. Three recent studies have provided support for this view by demonstrating a striking negative correlation between a protein's conformational stability and its propensity to aggregate in IBs [37,39,40]. This correlation might also exist for amyloids, as most mutations in genes encoding globular proteins that lead to deposition diseases destabilize the cooperatively folded conformation [41]. Thus, prokaryotes could be useful systems in which to study the conformational determinants of *in vivo* amyloid formation.

Amyloidogenic proteins in bacteria

Except for small peptides, the characterization of amyloid proteins usually requires their recombinant production. The similarity between amyloids and IBs explains why most pathogenic amyloid proteins accumulate in IBs when they are expressed in bacteria, regardless of whether they are unstructured or globular in the native conformation (Table 2). Notably, engineered variants that have reduced amyloidogenicity are invariably more soluble in bacteria than wild-type proteins [42], whereas an increased propensity to form amyloid promotes deposition into IBs [43]. Moreover, proteins designed to assemble into fibrils accumulate as IBs, whereas mutations that convert these proteins into monomeric β-sheets allow the proteins to remain soluble in the bacterial cytoplasm [30,44], highlighting how the determinants responsible for amyloid and IB aggregation overlap.

Cytotoxicity and infectivity of IBs

The mechanism by which amyloid structures exert their cytotoxic action remains unclear, but accumulating evidence points to prefibrillar assemblies being the pathogenic species. The severity of many deposition diseases correlates more closely with the levels of soluble oligomers than with the amount of fibrillar material. A recent report showed that the prevention of Aβ oligomerization is a valid therapeutic method for lessening or halting AD neurodegeneration. Amyloid oligomers, and specifically those isolated from AD brains, are resistant to sodium dodecyl sulfate (SDS) and thus are fairly stable [45]. Importantly, such SDS-stable oligomers also accumulate in bacterial cells shortly after the induction of Aβ expression, suggesting that amyloid fibrils and IBs containing pathogenic proteins share similar molecular pathways leading to their aggregation [7]. Prefibrillar aggregates of proteins that are unrelated to amyloidoses can be toxic, suggesting that a general mechanism underlies the cytotoxicity of unrelated oligomers [46]. If this is also true in bacteria, there will be important biotechnological consequences for the many polypeptides that accumulate in IBs, as soluble cytotoxic species might be co-purified with the target protein.

What about IBs? Are they toxic or protective? It has been argued that IBs are not harmful but rather have a detoxification role in bacteria. The same function has been proposed for the mature fibrils of several human pathogenic proteins, as they are devoid of any toxicity [47]. In fact, because all organisms face protein misfolding and deposition challenges, it seems reasonable that evolutionary strategies developed to reduce the harmful effects of cytotoxic assemblies. These mechanisms would act by sequestering sticky, partially folded species into regular supramolecular structures through specific interactions [1]. Accordingly, misfolded forms and soluble aggregates of recombinant proteins, but not large insoluble aggregates, can induce rearrangement of membrane lipids and of specific host proteins [48]. Nonetheless, it has been proposed that bacterial aggregates are not fully mature species but rather protofibrillar structures that have not evolved into highly ordered fibrils owing to kinetic and steric impediments [6,8]. This prefibrillar arrangement might account for the toxicity of IBs when administered to cultured mammalian cells [49]. Interestingly, the toxicity of the aggregates inversely correlates with the presence of ordered cross-β structure, indicating that, as for amyloids, loosely packed conformations constitute the most harmful species [46]. Also, IBs accumulate upon cell division in cells harbouring older poles. This effect is associated with a significant loss of productive ability (aging), relative to new-pole progeny, which is devoid of IBs [50]. This finding suggests that IBs are also toxic in bacteria and, accordingly, dividing cells segregate the damage, thereby promoting the likelihood that the population will perpetuate.

Prions represent a particular subclass of amyloids for which the aggregation process becomes self-perpetuating and infectious *in vivo*. Interestingly, the prion domain of *Saccharomyces cerevisiae* [PSI$^+$] forms fibrillar intracellular structures when expressed in *E. coli*. These aggregates are highly toxic for the bacteria, whereas prion variants that harbour amino acid substitutions that prevent the conformational transition toward the aggregated state become innocuous [51]. In addition, mouse prion protein (PrP) and fragments thereof also form amyloid-like IBs [8]. Because it is likely that human, cow or sheep recombinant prions would also form such structures, they should be used with caution, as their potential to infect humans must not be disregarded.

Concluding remarks and future perspectives

The detection of toxic amyloid-like conformations in bacteria raises unexpected safety concerns related to the biomedical and biotechnological uses of recombinant proteins. Nonetheless, because amyloids and IBs seem to share properties related to sequence, conformation and function, the ease with which bacteria can be genetically and biochemically manipulated allows their use as tools for performing chemical and genetic screens for inhibitors of protein aggregation. In a proof-of-principle experiment, Kim and coworkers [52] used an Aβ–GFP fusion to screen a chemical library for compounds that consistently abrogate aggregation in bacteria and amyloid fibril formation. In this assay, the solubility and/or aggregation behaviour of Aβ is coupled to the fluorescence of GFP. The selected compounds were shown to be *bona fide* inhibitors of Aβ fibril formation. These results pave the way for the development of novel, bacteria-based approaches for identifying agents that interfere with the earliest steps of amyloid aggregation. However, their utility for the study of conformational disorders in which the pathogenic agent is not a short peptide but instead a structurally more complex globular protein requires further investigation. Our knowledge of amyloids is likely to prove useful in the rational engineering of protein aggregation in cell factories, allowing us to expand the number of soluble polypeptides that are available for biomedical and biotechnological applications. To advance toward this objective we need effective ways to integrate accurate predictions of aggregation propensity, conformational stability and translation rates for each target protein. Finally, it remains to be explored whether intracellular bacterial proteins also assemble into amyloid-like structures under normal physiological conditions or whether this phenomenon only occurs during the overexpression of heterologous genes.

Acknowledgements

Work in our laboratory has been supported by grants BIO2007-68046 (Ministerio de Educación y Ciencia, Spain) and 2005-SGR00037 (Generalitat de Catalunya).

References

1 Chiti, F. and Dobson, C.M. (2006) Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.* 75, 333–366

2 Ventura, S. and Villaverde, A. (2006) Protein quality in bacterial inclusion bodies. *Trends Biotechnol.* 24, 179–185

3 Fernandez-Busquets, X. *et al.* (2008) Recent structural and computational insights into conformational diseases. *Curr. Med. Chem.* 15, 1336–1349

4 Bowden, G.A. *et al.* (1991) Structure and morphology of protein inclusion bodies in *Escherichia coli*. *Biotechnology (N. Y.)* 9, 725–730

5 Carrio, M. *et al.* (2005) Amyloid-like properties of bacterial inclusion bodies. *J. Mol. Biol.* 347, 1025–1037

6 Schrodel, A. and de Marco, A. (2005) Characterization of the aggregates formed during recombinant protein expression in bacteria. *BMC Biochem.* 6, 10

7 Morell, M. *et al.* (2008) Inclusion bodies: specificity in their aggregation process and amyloid-like structure. *Biochim. Biophys. Acta* 1783, 1815–1825

8 Wang, L. *et al.* (2008) Bacterial inclusion bodies contain amyloid-like structure. *PLoS Biol.* 6, e195

9 Jahn, T.R. and Radford, S.E. (2008) Folding versus aggregation: polypeptide conformations on competing pathways. *Arch. Biochem. Biophys.* 469, 100–117

10 Jahn, T.R. and Radford, S.E. (2005) The yin and yang of protein folding. *FEBS J.* 272, 5962–5970

11 Carrio, M.M. and Villaverde, A. (2005) Localization of chaperones DnaK and GroEL in bacterial inclusion bodies. *J. Bacteriol.* 187, 3599–3601

12 Doglia, S.M. *et al.* (2008) Fourier transform infrared spectroscopy analysis of the conformational quality of recombinant proteins within inclusion bodies. *Biotechnol. J.* 3, 193–201

13 de Groot, N.S. and Ventura, S. (2006) Protein activity in bacterial inclusion bodies correlates with predicted aggregation rates. *J. Biotechnol.* 125, 110–113

14 Garcia-Fruitos, E. *et al.* (2005) Aggregation as bacterial inclusion bodies does not imply inactivation of enzymes and fluorescent proteins. *Microb. Cell Fact.* 4, 27

15 Ami, D. (2005) Kinetics of inclusion body formation studied in intact cells by FT-IR spectroscopy. *FEBS Lett.* 579, 3433–3436

16 Morgan, G.J. *et al.* (2008) Exclusion of the native alpha-helix from the amyloid fibrils of a mixed alpha/beta protein. *J. Mol. Biol.* 375, 487–498

17 Glover, J.R. *et al.* (1997) Self-seeded fibers formed by Sup35, the protein determinant of [PSI$^+$], a heritable prion-like factor of *S. cerevisiae*. *Cell* 89, 811–819

18 Taylor, G. *et al.* (1986) Size and density of protein inclusion bodies. *Biotechnology* 4, 553–557

19 Balguerie, A. *et al.* (2003) Domain organization and structure-function relationship of the HET-s prion protein of *Podospora anserina*. *EMBO J.* 22, 2071–2081

20 Jarrett, J.T. and Lansbury, P.T., Jr (1993) Seeding 'one-dimensional crystallization' of amyloid: a pathogenic mechanism in Alzheimer's disease and scrapie? *Cell* 73, 1055–1058

21 Krebs, M.R. *et al.* (2004) Observation of sequence specificity in the seeding of protein amyloid fibrils. *Protein Sci.* 13, 1933–1938

22 Jones, E.M. and Surewicz, W.K. (2005) Fibril conformation as the basis of species- and strain-dependent seeding specificity of mammalian prion amyloids. *Cell* 121, 63–72

23 de Groot, N.S. *et al.* (2008) Studies on bacterial inclusion bodies. *Future Microbiol.* 3, 423–435

24 Carrio, M.M. *et al.* (1998) Dynamics of *in vivo* protein aggregation: building inclusion bodies in recombinant bacteria. *FEMS Microbiol. Lett.* 169, 9–15

25 Ventura, S. *et al.* (2004) Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case. *Proc. Natl. Acad. Sci. U. S. A.* 101, 7258–7263

26 Fernandez-Escamilla, A.M. *et al.* (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.* 22, 1302–1306

27 Tartaglia, G.G. and Vendruscolo, M. (2008) The Zyggregator method for predicting protein aggregation propensities. *Chem. Soc. Rev.* 37, 1395–1401

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

29 Morimoto, A. *et al.* (2004) Analysis of the secondary structure of beta-amyloid (Aβ42) fibrils by systematic proline replacement. *J. Biol. Chem.* 279, 52781–52788

30 West, M.W. *et al.* (1999) *De novo* amyloid proteins from designed combinatorial libraries. *Proc. Natl. Acad. Sci. U. S. A.* 96, 11211–11216

31 Idicula-Thomas, S. and Balaji, P.V. (2005) Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in *Escherichia coli*. *Protein Sci.* 14, 582–592

32 Tartaglia, G.G. *et al.* (2007) Life on the edge: a link between gene expression levels and aggregation rates of human proteins. *Trends Biochem. Sci.* 32, 204–206

33 Hoffmann, F. and Rinas, U. (2004) Roles of heat-shock chaperones in the production of recombinant proteins in *Escherichia coli*. *Adv. Biochem. Eng. Biotechnol.* 89, 143–161

34 Espargaro, A. *et al.* (2008) Kinetic and thermodynamic stability of bacterial intracellular aggregates. *FEBS Lett.* 582, 3669–3673

35 Uversky, V.N. (2008) Amyloidogenesis of natively unfolded proteins. *Curr. Alzheimer Res.* 5, 260–287

36 Klein, J. and Dhurjati, P. (1995) Protein aggregation kinetics in an *Escherichia coli* strain overexpressing a *Salmonella typhimurium CheY* mutant gene. *Appl. Environ. Microbiol.* 61, 1220–1225

37 Espargaro, A. *et al.* (2008) The *in vivo* and *in vitro* aggregation properties of globular proteins correlate with their conformational stability: the SH3 case. *J. Mol. Biol.* 378, 1116–1131

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

39 Calloni, G. *et al.* (2005) Investigating the effects of mutations on protein aggregation in the cell. *J. Biol. Chem.* 280, 10607–10613

40 Mayer, S. *et al.* (2007) Correlation of levels of folded recombinant p53 in *Escherichia coli* with thermodynamic stability *in vitro*. *J. Mol. Biol.* 372, 268–276

41 Canet, D. *et al.* (2002) Local cooperativity in the unfolding of an amyloidogenic variant of human lysozyme. *Nat. Struct. Biol.* 9, 308–315

42 Wigley, W.C. *et al.* (2001) Protein solubility and folding monitored *in vivo* by structural complementation of a genetic marker protein. *Nat. Biotechnol.* 19, 131–136

43 Hammarstrom, P. *et al.* (2003) D18G transthyretin is monomeric, aggregation prone, and not detectable in plasma and cerebrospinal fluid: a prescription for central nervous system amyloidosis? *Biochemistry* 42, 6656–6663

44 Wang, W. and Hecht, M.H. (2002) Rationally designed mutations convert *de novo* amyloid-like fibrils into monomeric beta-sheet proteins. *Proc. Natl. Acad. Sci. U. S. A.* 99, 2760–2765

45 Walsh, D.M. *et al.* (2002) Amyloid-beta oligomers: their production, toxicity and therapeutic inhibition. *Biochem. Soc. Trans.* 30, 552–557

46 Bucciantini, M. *et al.* (2002) Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature* 416, 507–511

47 Merlini, G. and Bellotti, V. (2003) Molecular mechanisms of amyloidosis. *N. Engl. J. Med.* 349, 583–596

48 Ami, D. *et al.* (2009) Effects of recombinant protein misfolding and aggregation on bacterial membranes. *Biochim. Biophys. Acta* 1794, 263–269

49 Gonzalez-Montalban, N. *et al.* (2007) Amyloid-linked cellular toxicity triggered by bacterial inclusion bodies. *Biochem. Biophys. Res. Commun.* 355, 637–642

50 Lindner, A.B. *et al.* (2008) Asymmetric segregation of protein aggregates is associated with cellular aging and rejuvenation. *Proc. Natl. Acad. Sci. U. S. A.* 105, 3076–3081

51 Ono, B. *et al.* (2008) Effects of mutations in yeast prion [PSI$^+$] on amyloid toxicity manifested in *Escherichia coli* strain BL21. *Prion* 2, 37–41

52 Kim, W. *et al.* (2006) A high-throughput screen for compounds that inhibit aggregation of the Alzheimer's peptide. *ACS Chem. Biol.* 1, 461–469

53 Hart, R.A. *et al.* (1990) Protein composition of *Vitreoscilla* hemoglobin inclusion bodies produced in *Escherichia coli*. *J. Biol. Chem.* 265, 12728–12733

54 Monsellier, E. *et al.* (2007) The distribution of residues in a polypeptide sequence is a determinant of aggregation optimized by evolution. *Biophys. J.* 93, 4382–4391

55 Shorter, J. and Lindquist, S. (2008) Hsp104, Hsp70 and Hsp40 interplay regulates formation, growth and elimination of Sup35 prions. *EMBO J.* 27, 2712–2724

56 Makin, O.S. and Serpell, L.C. (2005) X-ray diffraction studies of amyloid structure. *Methods Mol. Biol.* 299, 67–80

57 Tycko, R. (2006) Solid-state NMR as a probe of amyloid structure. *Protein Pept. Lett.* 13, 229–234

58 Barnham, K.J. *et al.* (2006) Delineating common molecular mechanisms in Alzheimer's and prion diseases. *Trends Biochem. Sci.* 31, 465–472

59 Nagy, Z. (2000) Cell cycle regulatory failure in neurones: causes and consequences. *Neurobiol. Aging* 21, 761–769

60 Singh, D. *et al.* (2007) Mixed oligomer formation between human αA-crystallin and its cataract-causing G98R mutant: structural, stability and functional differences. *J. Mol. Biol.* 373, 1293–1304

61 Lopes, D.H. *et al.* (2004) Amyloidogenicity and cytotoxicity of recombinant mature human islet amyloid polypeptide (rhIAPP). *J. Biol. Chem.* 279, 42803–42810

62 Wang, J. *et al.* (2003) Overexpression and purification of recombinant atrial natriuretic peptide using hybrid fusion protein REF–ANP in *Escherichia coli*. *Protein Expr. Purif.* 28, 49–56

63 Ding, M.S. *et al.* (2005) *Sheng Wu Gong Cheng Xue Bao* 21, 198–203

64 Umetsu, M. *et al.* (2005) Nondenaturing solubilization of β2 microglobulin from inclusion bodies by L-arginine. *Biochem. Biophys. Res. Commun.* 328, 189–197

65 Ishikawa, H. *et al.* (1999) Large-scale preparation of recombinant human calcitonin from a multimeric fusion protein produced in *Escherichia coli*. *J. Biosci. Bioeng.* 87, 296–301

66 Gerhartz, B. *et al.* (1998) Two stable unfolding intermediates of the disease-causing L68Q variant of human cystatin C. *Biochemistry* 37, 17309–17317

67 Swietnicki, W. *et al.* (1998) Familial mutations and the thermodynamic stability of the recombinant human prion protein. *J. Biol. Chem.* 273, 31048–31052

68 Nagao, Y. *et al.* (2000) DMSO and glycerol reduce bacterial death induced by expression of truncated N-terminal huntingtin with expanded polyglutamine tracts. *Biochim. Biophys. Acta* 1502, 247–256

69 Helms, L.R. and Wetzel, R. (1996) Specificity of abnormal assembly in immunoglobulin light chain deposition disease and amyloidosis. *J. Mol. Biol.* 257, 77–86

70 Redwan, El-RM. *et al.* (2008) Synthesis of the human insulin gene: protein expression, scaling up and bioactivity. *Prep. Biochem. Biotechnol.* 38, 24–39

71 Paladini, R.D. *et al.* (1995) cDNA cloning and bacterial expression of the human type I keratin 16. *Biochem. Biophys. Res. Commun.* 215, 517–523

72 Leinweber, B. *et al.* (2004) Aggregation of ALS mutant superoxide dismutase expressed in *Escherichia coli*. *Free Radic. Biol. Med.* 36, 911–918

73 Li, M. and aand Su, Z. (2002) Refolding human lysozyme produced as an inclusion body by urea concentration and pH gradient ion exchange chromatography. *Chromatographia* 56, 33–38

74 Rajan, R.S. *et al.* (2001) Specificity in intracellular protein aggregation and inclusion body formation. *Proc. Natl. Acad. Sci. U. S. A.* 98, 13060–13065

75 Milewski, M.I. *et al.* (2002) Aggregation of misfolded proteins can be a selective process dependent upon peptide composition. *J. Biol. Chem.* 277, 34462–34470

Protein Aggregation Profile of the Bacterial Cytosol

Natalia S. de Groot[1] and Salvador Ventura[1]*

[1]Departament de Bioquímica i Biologia Molecular and Institut de Biotecnologia i Biomedicina and
Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain

Running title: Aggregation propensity of *E. coli* cytosol

*Address correspondence to: Salvador Ventura (Tel. +34-935868147; Fax +34 935811264; Email: salvador.ventura@uab.es)

Abstract

Background: Protein misfolding is usually deleterious for the cell, either as a consequence of the loss of protein function or the build up of insoluble and toxic aggregates. The aggregation behaviour of a given polypeptide is strongly influenced by the intrinsic properties encoded in its sequence. This has allowed the development of effective computational methods to predict protein aggregation propensity.

Methodology/Principal Findings: Here, we use the AGGRESCAN algorithm to approximate the aggregation profile of an experimental cytosolic Escherichia coli proteome. The analysis indicates that the aggregation propensity of bacterial proteins is associated with their length, conformation, location, function and abundance. The data are consistent with the predictions of other algorithms on different theoretical proteomes.

Conclusions/Significance: Overall, the study suggests that the avoidance of protein aggregation in functional environments acts as a strong evolutionary constraint on polypeptide sequences in both prokaryotic and eukaryotic organisms.

Introduction

In the cellular context, it is the native protein fold that determines the biological function. Therefore, protein misfolding is usually associated with the impairment of essential cellular processes. In many cases, the assembly of misfolded polypeptides into cytotoxic aggregates mediates this deleterious effect. Accordingly, protein deposition is linked to the onset of more than 40 different human disorders [1]. In these diseases, proteins usually self-assemble into highly ordered, b-sheet enriched, supramolecular structures known as amyloid fibrils. However, the aggregation into amyloid conformations is not restricted to disease-related proteins but appears to be a generic property of polypeptides [2,3,4]. Moreover, although traditionally thought to be restricted to eukaryotic cells, recent studies provide compelling evidence for the formation of toxic amyloid assemblies inside bacteria [5,6,7,8]. In this scenario, because all organisms face the important challenges of protein misfolding and aggregation, the existence of evolutionarily conserved strategies to avoid the deleterious effects of undesired protein deposition is likely.

The main intrinsic properties that determine protein aggregation have been defined and different computational approximations [9,10,11,12,13,14,15,16,17,18,19,20] have exploited them to predict with reasonable accuracy the regions of proteins with the highest aggregation propensity, also called Hot Spots (HS), as well as the overall protein aggregation propensity. Most of these algorithms only require the protein primary sequence as the input, allowing their easy implementation for the large-scale analysis of protein sets [1,21,22,23,24,25,26,27]. Rosseau and co-workers used the TANGO algorithm to analyse the aggregation propensity of 28 complete proteomes, finding that polypeptides without a defined structure, and therefore with a solvent-accessible sequence, are less aggregation-prone than globular proteins [27]. The same group demonstrated that in *Escherichia coli* (*E. coli*), there is a bias towards the presence of residues with a low aggregation propensity flanking aggregation-prone stretches and that chaperones seem to have evolved to recognise these sequence features [27]. Tartaglia and co-workers employed their algorithm to compare the deposition tendency of different eukaryotic proteomes. They observed that the proteins of higher eukaryotes, and specifically of those with a longer lifespan, tend to be less aggregation-prone [24]. Moreover, the study of the Saccharomyces cerevisiae proteome revealed that in this organism, the protein aggregation propensity is associated to both protein function and localisation [23]. More recently, Chiti and co-workers used the Zyggregator program to analyse the aggregation tendency of the human proteome, their results recapitulated those of the above-discussed studies and additionally showed that long human proteins posses less-intense aggregation peaks than shorter ones [21].

Overall, the above-mentioned studies have provided important insights on how organisms deal with protein aggregation. However, all of them have used theoretical proteomes derived from the predicted ORFs in the different genomes as protein data sets. This means that they do not provide selective information on the specific aggregation propensities of the real set of proteins that are present in a given cell under specific conditions and, perhaps more importantly, on how aggregation correlates with the real abundance of a protein in a cell. To address this issue, we have used AGGRESCAN, an algorithm previously developed by our group [10,28], to analyse the aggregation propensity of the experimentally determined cytosolic proteome of the *E. coli* strain MC4100. This protein set comprises more than 1000 different proteins for which the individual abundance in the cytoplasmic fraction could be experimentally measured [29]. The results of our analyses provide new insights into the relationship between the intrinsic deposition propensities, cellular protein concentrations and protein expression regulation. In addition, the data recapitulate most of the previous observations on virtual proteomes. The overall analysis suggests that natural selection modulates proteins aggregation propensities according to their cellular function, structure, concentration and localization.

Results and Discussion

AGGRESCAN parameters and the protein data set

AGGRESCAN appears to be one of the best suited algorithms to analyse the aggregation propensities of bacterial cytoplasmic proteins because it is based on an aggregation-propensity scale for natural amino acids derived from the experimental analysis of the properties of an aggregation-prone polypeptide when expressed in the *E. coli* cytoplasm [25,30,31]. From the different outputs provided by the program, in the present work we have selected the following parameters: the number of Hot Spots (HS) in a sequence (NnHS), the total area of these aggregation-prone regions (THSAr) and the global protein aggregation propensity (Na4vSS). We choose this particular set of values because, in AGGRESCAN, all of them are normalized relative to the number of amino acids in the sequence, allowing the direct comparison of proteins with different sizes (Figure 1).

The protein data set includes 1103 different proteins whose presence could be experimentally detected in the bacterial cytosol [29]. From these, PSORT [32,33] classified 579 as cytoplasmic, 334 as having an unknown location and 190 as belonging to other subcellular compartments. The first and second subsets were joined and considered as cytosolic proteins, with the exception of those proteins for which the experimental evidence suggested that they were not mainly located in the cytoplasm (49). In addition, the proteins assigned by PSORT to other compartments but experimentally shown to be cytoplasmic (11) were also included in the cytosolic set, resulting in a total of 875 polypeptides. This subset was used for all of the subsequent analyses, except for the calculation of the aggregation propensities of bacterial compartments, where the whole data set was employed. AGGRESCAN was run and the above-mentioned values were calculated for each protein in the set.

The cytosolic proteins abundance correlate with their aggregation propensity

Most protein aggregation processes follow a nucleation-polymerization scheme, in which the formation of the initial aggregation nuclei represents the rate-limiting step of the overall process. Nucleation processes correspond to second-order reactions and therefore the rate of protein aggregation is strongly dependent on the initial protein concentration. Therefore, the effective intracellular concentration becomes an important parameter when studying protein aggregation *in vivo*. The number of mRNAs in the bacterial cytosol encoding a given protein can vary from 1 to 100,000 [34]. Ishihama and co-workers developed the exponentially modified Protein Abundance Index (emPAI) to approximate the real concentration of a protein in a living cell. This index associates the number of mass spectrometrysequenced peptides for each experimentally detected protein with its concentration in a given preparation. Later on, they applied this approach to successfully calculate the abundance of individual proteins in the bacterial cytosolic fraction [29,35]. We used these data to compare the aggregation features of the 10% most abundant cytosolic proteins (MAP) with those of the 10% least abundant ones (LAP).

The normalized average number of HS (NnHS) is approximately three in both groups. However, sequences devoid of any HS were observed only in the MAP group and sequences with NnHS values ≤ 2 were also more frequent in this subset (Figure 2A). Nevertheless, the frequency of proteins with NnHS values ≥ 5 was also higher in this group. The graphic of the THSAr closely resembles that of the NnHS, indicating that no important differences exist in the area associated with the aggregation-prone regions between the two groups (Figure 2B). In contrast, the overall aggregation propensity of LAP sequences is clearly much higher than that of MAP (Figure 2C). To study the degree of association between the abundance of cytosolic proteins and their overall aggregation propensity, the protein set was divided in 50 groups according to their abundance. The average Na4vSS value of each group was calculated and the two parameters were compared (Figure 2D). A significant correlation was observed (R=0.71), indicating a relationship between the polypeptide solubility and the abundance levels in the cytosol. This correlation suggests an evolutionary selection of bacterial cytoplasmic proteins to minimize their deposition at the concentrations required for their proper biological functions. The higher solubility of the MAP would work to prevent the aggregation of these proteins even if they become concentrated at specific sub-cytosolic locations. Moreover, because of their high concentrations, their low deposition propensity would contribute significantly to decrease the overall cytosol aggregation tendency and prevent the initiation of spontaneous, non-specific aggregation processes that can deplete the cell of less represented and/or functionally important proteins.

The results suggest that the MAP would be less aggregation-susceptible than the LAP not because they have fewer or weaker aggregation-prone regions, but because these segments are located in a much more soluble sequence context, which counteracts their self-assembly tendency. Therefore, we analysed whether MAP and LAP sequences differed in their amino acid composition (Figure 3A). One of the most striking differences between the compositions of the two protein sets is a strong bias for a higher presence of Lys residues in the MAP set. Also, Glu is more represented in the MAP set, but the difference compared to the LAP set is lower than in the case of Lys. The other charged residues, Arg and Asp, are found in similar amounts in both protein sets. This causes the overall theoretical pI of the MAP set (8.48) to be higher than that of the LAP set (6.71). The pH of the *E. coli* cytosol is thought to be around 7.5 [36]. Accordingly, the overall deviation from the physiological pH is higher for the MAP set (+ 0.98 units) than for the LAP set (-0.79 units). We analysed the individual contributions of polypeptides to these deviations by measuring the percentage of proteins whose pI deviated two pH units below or above the physiological pH. According to this criterion, highly acidic and basic polypeptides constituted 27% and 53% of the MAP set, respectively; in contrast, to 20% and 10% in the LAP set. This means that, as a general trend, proteins in the LAP group have a pI closer to the cytosolic pH than those in the MAP set. To test whether there is any relationship between the theoretical pI of a protein and its predicted deposition propensity, we grouped the polypeptides in the cytosolic fraction according to their pIs. Then the average Na4vSS was calculated for each group and plotted against the pI. The resulting graphic shows that proteins with a pI distant from the bacterial cytosolic pH, either more acidic or more basic, have lower aggregation propensities (Figure 3B), explaining why MAP tend to populate the extremes of the pI distribution. Because the net charge of a protein at a given pH depends on its pI, these results are in excellent agreement with previous observations indicating that, *in vitro,* the net charge of a protein anti-correlates with its aggregation propensity [37,38,39].

The abundance of both acidic and basic proteins in the MAP set can be attributed to the overrepresentation of Glu and especially Lys residues and suggests that these excesses of charged residues do not mutually compensate for each other in MAP. Importantly, Lys is by far the least frequently buried residue among the 20 natural amino acids [40]. This is because it needs two other residues to hydrogen bond to its side chain nitrogen atom when it is located in the core of the protein. Glu residues are also less frequently buried in the core than Asp because they have a weaker tendency to bond to the local main chain. This suggests that in MAP, these residues are preferentially located at the surface in the folded conformation. Interestingly enough, it has been recently shown that increasing the net charge in the surface of a globular protein is a very effective strategy to prevent its aggregation, even in harsh conditions [41,42]. It is likely that the *E. coli* cytosol would exploit the same strategy to prevent the aggregation of highly abundant polypeptides.

Apart from the charge, another property that strongly influences the overall aggregation propensity of a protein sequence is its hydrophobicity [2,25,43]. Interestingly, the proportion of hydrophobic residues in these two groups is not dramatically different: 41.6% and 42.4% for MAP and LAP, respectively. However, a bias toward the presence of larger residues, like Trp or Tyr, in the place of smaller residues, like Val, is observed in LAP (Figure 3A). This suggests that LAP could be overall more hydrophobic than MAP. We used the grand average of the hydropathicity (GRAVY) as measure of the hydrophobicity of both protein sets [44]. The average GRAVY scores are -0.24 and -0.36 for LAP and MAP, respectively. Also, 38% of MAP have a GRAVY value below -0.5, in contrast with only 10% of LAP. Both data indicate that MAP tend to be less hydrophobic than LAP. This is likely because hydrophobicity is strongly associated with the aggregation propensity, as shown when analyzing the correlation between these two parameters in the complete cytosolic set (R=0.88) (Figure 3C). It is worth mentioning that Cys residues are underrepresented in both cytosolic protein sets, but especially in MAP, relative to the conjunct of natural proteins. Reducing conditions prevail in the cytoplasm and disulfide bonds do not normally form correctly in this compartment, which can result in the accumulation of misfolded and inactive proteins [45]. The low content of Cys in bacterial cytosolic proteins is likely the result of a negative selection to avoid these phenomena.

The correlation between the effective protein concentration and aggregation propensity suggests that this relationship is controlled at the gene level, providing the cell with the versatility and adaptability necessary to react to different environmental conditions and/or cellular states. The codon usage can be employed to approximate the protein abundance, obtaining similar estimations to those derived from quantifying mRNA expression levels [46,47]. We used the codon adaptation index (CAI) as a measure of the codon usage. Low CAI values are associated with low expression levels and high CAI values correspond to high expression levels [29]. The comparison of the 10% of genes encoding cytoplasmic proteins with the higher and lower CAIs shows that both sets present distinctive aggregation features. The low CAI group presents higher Na4vSS values than the highly expressed one (Figure 4A). In addition, when all the cytoplasmic proteins are arranged into 20 groups according to their CAI values, a significant correlation between this parameter and the protein aggregation propensity (R=0.77) is observed (Figure 4B).

These results are in agreement with those obtained using EMPAI as a measure of the experimental protein concentration, which overall suggests that the relationship between the protein concentration and aggregation propensity is controlled at the gene expression level. Confirming this hypothesis, a relationship between the mRNA expression levels and protein solubility in *E. coli* has been recently described [48]. Beginning with the AGGRESCAN scale, Tartaglia and co-workers also observed that sequences with the highest mRNA expression levels are less aggregation-prone and *vice versa*. We have previously shown that recombinant soluble proteins have, on average, lower aggregation propensities than those that accumulate as insoluble deposits in the bacterial cytosol upon heterologous overexpression [10]. Extending this observation, Tartaglia and co-workers were able to theoretically forecast the solubility of recombinant proteins in bacteria from their expected expression levels [48]. These data converge to indicate that successfully expressed recombinant proteins would resemble the MAP more than the LAP. The sum of the squared differences between the amino acid composition of a set of soluble recombinant proteins [10] and that of the MAP and LAP groups is 79.5 and 114.9, respectively, thus providing support for this hypothesis.

A relationship between the molecular weight and aggregation propensity

Chiti and co-workers have recently suggested that long human protein sequences have been shaped by evolution in order to reduce their intrinsic aggregation properties [21]. To study the relationship between the protein size and deposition propensity in bacterial cytosolic proteins, we grouped proteins into 50 sets according to their molecular weights (MW) and the average Na4vSS for each particular group was calculated. As shown in Figure 5A, the nature of the relationship between the aggregation propensity and protein length depends on the particular size of the polypeptide. For small proteins, up to approximately 20 kDa in size, the increase in MW is associated with a rapid increase in the aggregation propensity (R=0.92). Once this size limit is over-passed, the correlation is inverted and further increases in size are linked to a predicted slow, but progressive, increase in solubility (R=0.75). If we consider the shape of a protein close to a sphere, then its surface area would be approximately proportional to the two-thirds power law of its volume [49]. This implies that, for globular proteins, the relative size of the core grows with protein size [50]. Because hydrophobic residues usually occupy the core of the protein to avoid interaction with water molecules, it is deduced that the proportion of hydrophobic residues, and therefore the overall aggregation propensity, increases with the protein size. Nevertheless, in real proteins, the correlation between the protein size and the fraction of hydrophobic amino acids appears to apply only for proteins until 170 residues [40], in agreement with the observation that the aggregation propensity attains maximum values in this size range. The protein aggregation propensity might act as a determinant of protein size and could be the underlying reason explaining why, above the ~20 kDa limit, the ratio between hydrophobic and hydrophilic residues does not increase significantly with size [51,52]. An important implication of the volume/surface relationship in globular proteins, is that, if the proportion of hydrophobic residues is approximately constant, the number of polar residues buried inside the structure should increase with protein size [51,53,54]. Because charged residues are more hardly accommodated inside proteins than other polar residues, long proteins tend to have fewer charges [55], which together with their slow folding rates [56], would make these proteins aggregation susceptible. According to our data, in *E. coli* polypeptides, these effects are partially compensated by an overall decreased sequence aggregation propensity. Importantly, above the 20-kDa limit, the NnHS values steadily decrease with the protein size indicating that in longer proteins (Figure 5B), the HS tend to be more distant in the sequence. Interestingly enough, the main bacterial chaperones, GroEL and DnaK, interact poorly with proteins smaller than 20 kDa and display a preference for larger substrates (Figure 5A) [57,58,59], suggesting the presence of redundant mechanisms to reduce the aggregation propensity of long bacterial proteins, as previously described for the human proteome [21].

The composition of hot spot and gatekeeper stretches

It has been suggested that evolution exploits negative design principles to modulate protein deposition by placing residues that counteract aggregation at the flanks of HS [21,27,60]. These residues would act as gatekeepers [27] and reduce the protein propensity to self-assemble into macromolecular aggregates. At the same time, it appears that the cellular quality control has evolved to recognize and block these sequence patterns [21,27]. Accordingly, several disease-associated mutations have been linked to the disruption of gatekeeper stretches [61]. To confirm these observations, we proceeded to study whether, in bacterial cytosolic proteins, HS and their flanking sequence stretches differ in composition (Figure 6A). The comparison of the amino acid frequency in the these regions with their natural abundance shows that hydrophobic and aggregation-promoting residues (Val, Phe, Ile, Tyr Met and Leu) are overrepresented inside HS and, on the contrary, that flanking regions are enriched with polar and soluble residues (Arg, Asp, Glu, Asn Lys and Gln). The rate between the frequency of each amino acid inside the HS and at the flanks evidenced that Phe displays a high preference for being a component of aggregation-prone regions (Figure 6B). In contrast, the charged Arg, Lys, Asp and Glu residues display a high preference for being at the flanks (Figure 6C). The gatekeeper action of these residues is exerted through the repulsive effect of the charge (Arg, Lys, Asp and Glu) and the increase in entropy penalties upon assembly (Arg and Lys). Our data are in excellent agreement with the distribution found using the TANGO and Zygregator algorithms on the theoretical *E. coli* and human proteomes [21,27], indicating that the protective action of the flanking residues acts on the combination of proteins that are being effectively expressed in the bacterial cytosol. As described above, another important gatekeeper residue is Pro, which acts as a beta-breaker. Because AGGRESCAN considers the presence of a Pro residue in a sequence stretch incompatible with this sequence being a HS, its frequency could not calculated.

The relationship between the aggregation propensity and protein function in cytosolic proteins

The set of genes in an operon share a common gene expression regulation and are generally connected by their biological function. As a result, proteins encoded by the same operon are suggested to be present in similar amounts in the cell [29]. The observed association between protein aggregation and abundance would imply that polypeptides in the same operon should have related aggregation propensities. In agreement with this hypothesis, the standard deviation of the Na4vSS value between proteins regulated by the same operon is lower in 78 % of the cases (25 of 33) than the standard deviation in the complete set of proteins (7,72 Na4vSS) that could be ascribed to a particular operon (Figure 7). This suggests again a link between protein aggregation propensities and the rates of transcriptional initiation.

The impact of protein aggregation on cellular function would be ultimately associated to individual fitness. Therefore, it is conceivable that evolution would select for an overall decreased aggregation propensity in operons performing essential cellular functions. To explore this possibility, the bacterial operons where divided in two groups according to their Na4vSS values, the lower (LA operons) and higher (HA operons) than the mean aggregation propensity of the complete operon protein set (-6.4 Na4vSS), respectively. The essentiality of approximately half of the proteins in each subset has been annotated via genetic footprinting or knockout experiments [62,63]. Importantly, considering only the annotated polypeptides, LA operons regulate 85% of essential proteins and 15% of nonessential ones. In contrast, HA operons encode a similar proportion of essential and nonessential proteins, 48% and 52% respectively (Table 1), suggesting that the sequences of essential bacterial cytoplasmic proteins suffer a stronger selection against deposition than those of nonessential ones, as previously proposed for different eukaryotic organisms [64].

A deeper analysis of the two operon subsets reveals that LA operons control the expression of 95% of the bacterial ribosomal proteins that could be ascribed to a given operon (Table 1). This suggests, because of their crucial function, ribosomal proteins might display differential aggregation traits. The analysis of the 53 ribosomal proteins detected in the cytosolic extract shows that these polypeptides display fewer HS and lower Na4vSS values than the rest of proteins in the bacterial cytoplasm (Figures 8A and 8B). Low aggregation propensities have been also predicted for human ribosomal proteins [21] suggesting a common evolutionary pressure for highly soluble ribosomal proteins. Ribosomal proteins are commonly characterised by the presence of unstructured sequence stretches. These regions act as "structural mortar". They have evolved to bind the ribosomal RNA and thereafter acquire a partial ordered structure that fills the gaps of the ribosome structure [65]. These unstructured regions might confer ribosomal proteins with a lower aggregation propensity than the rest of the cytosolic domains, in line with the idea that disordered sequences have been evolutionary selected to avoid the presence of aggregation-prone residues as a strategy to prevent the self-assembly of the fully solvent-exposed polypeptide chain in the absence of a protective secondary structure [22]. To confirm that this relationship applies for bacterial cytosolic proteins, we identified those polypeptides classified as intrinsically unstructured (IUP) according to the Disprot Database [66], calculated their aggregation parameters and compared them with the rest of cytosolic proteins (Figures 8C and 8D). As expected, bacterial cytosolic IUPs present a significantly decreased aggregation propensity. The difference in the aggregation propensity between the folded and disordered protein regions becomes even clearer if we only consider the fully unstructured sequences in IUPs and not the whole protein (Figures 8E and 8F).

Bacterial proteins in the periplasm and inner and outer membranes possess characteristic aggregation propensities

Eukaryotic cells consist of a complex collection of compartments characterised by different environmental conditions and molecular compositions [67,68]. It is suggested that proteins located in a particular eukaryotic subcellular location have been evolutionary selected to fold and avoid protein aggregation in this environment [21,22,23,24]. Bacterial proteins are found in other compartments apart from the cytosol, like the periplasm and the inner and outer membranes. Presumably their aggregation properties would be also adapted for their optimal function at those subcellular locations. As described above, the original data set used in the present work was enriched in cytoplasmic proteins but contained also polypeptides assigned to other cellular places. We took advantage of this protein diversity to analyse the aggregation properties of proteins residing in different compartments.

Cytoplasmic and periplasmic proteins exhibit a similar average aggregation propensity although a sharper distribution of Na4vSS values was observed in the periplasm, in which proteins with extreme aggregation propensities were absent. (Figure 9). The number of HS and their associated areas are lower in periplasmic proteins, suggesting that despite having a content of aggregation-prone residues similar to that of cytosolic proteins, these residues are differently arranged in the sequence (Figure 9). This is consistent with the observation that the average number of alternating hydrophobic/hydrophilic stretches (>5 residues) is 30% higher in periplasmic proteins, which might indicate a tendency to reduce the presence and impact of contiguous aggregation-prone regions. In line with this hypothesis, Chang and co-workers demonstrated experimentally that periplasmic proteins are preferentially resistant against aggregation under denaturing conditions and that this behaviour is not related to a higher thermodynamic stability, but rather to sequence characteristics [69]. This property can be evolutionary advantageous in the periplasm that, in contrast to the cytosol, lacks a sophisticated cellular system to control protein quality and avoid aggregation [68] and is separated from the outside solution by a highly permeable outer membrane that provides limited protection against environmental variations.

The gram-negative bacterial inner membrane (IM) is a semipermeable shield that preserves the cytoplasm environment. The proteins associated with the IM are principally composed of α-helices and could have a variable number of transmembrane segments (TS) per protein [67]. These regions are stable in the hydrophobic environment of this lipid bilayer due to a primary sequence rich in apolar residues. In this sense, it is necessary for a protein to have a stretch of 15-25 residues to transverse the membrane bilayer. Consequently, the extraction and analysis of these proteins in aqueous solvents frequently causes aggregation problems [67]. In agreement with these data, AGGRESCAN shows that IM proteins possess the highest aggregation propensities of all bacterial proteins (Figure 9C). Surprisingly, IM proteins contain a number of HS similar to that in cytoplasmic proteins (Figure 9D). However, in the IM, the area associated to these HS is much larger, indicating that they are significantly longer and/or contain more aggregation-prone residues (Figure 9E). These results are consistent with the observations obtained with TANGO, which also showed that membrane-associated proteins do not contain a higher amount of beta-aggregation nucleating regions than the proteins located in the cytoplasm [22]. Interestingly, when the Na4vSS values of the IM proteins were plotted as a dotted distribution, the existence of two IM protein groups become evident: a first group with an aggregation propensity similar to that of cytosolic proteins and a second group with particularly high Na4vSS values (Figure 10). We found that the main difference between these groups is the number of TS. The TMHMM version 2.0 [70] program calculated that 83% of the proteins in the first group contain fewer than three TS whereas 89% of the second group has more than three TS (Figure 9, Table 1). To decipher whether the different aggregation propensities exhibited by these two IM protein subsets was associated with particular biological functions (Figure 11), we consulted the functional descriptions collected in the Functional Catalogue Database (FunCatDB) [71] and in the Protein Knowledgebase (UniProtKB) [72,73]. According to the FunCatDB, IM proteins with high Na4vSS are preferably related to "transport facilitation" whereas functions like "cellular communication" or "protein fate" appear to be associated with IM with lower aggregation propensities. In agreement with these data, according to UniProtKB, IM proteins with a high aggregation propensity are preferentially involved in "electron transport" and "sugar transport" whereas IM proteins with low Na4vSS are associated to processes like "protein binding" and "ATP binding". Because, according to our analysis, IM with high aggregation propensities also contain many TS, they must be totally inserted in the membrane, limiting their actions to functions principally related to transport and respiratory activities. In contrast, polypeptides with low aggregation propensities are anchored in the IM by only one or two transmembrane helices, the rest of the protein being available to assume different biological activities like signal transduction [74].

Outer membrane (OM) proteins are thought to be located in a hydrophobic environment, and consequently, they are expected to have a high aggregation tendency. However, they exhibit a low aggregation propensity according to all AGGRESCAN parameters (Figure 9). In fact, the outer membrane acts as a permeable barrier to hydrophobic substances. In general, OM proteins display a beta barrel structure that encloses a hydrophilic cavity covered by a hydrophobic outer layer. The presence of an apolar hollow space is essential for their function as porins. Interestingly, this particular assembly is achieved by alternating hydrophobic and hydrophilic segments [75,76]. As a result, OM proteins display two times more alternating hydrophobic/hydrophilic stretches (>5 residues) than cytoplasmic proteins. The presence of these characteristic polar regions reduces the protein hydropathy and overall aggregation propensity but also limits the number and area associated with the HS. These properties could be important not only for their biological function but also for their biogenesis. As recently reviewed by Knowles and co-workers, the folding of proteins into the outer membrane presents important challenges to Gram-negative bacteria because they must migrate from the cytosol, through the inner membrane and into the periplasm before they could be recognized by the beta-barrel assembly machinery and inserted into the outer membrane [77]. In most of these steps and compartments, the protein is unfolded and accordingly sequences with reduced aggregation propensities would represent a selective advantage.

In the present study, we have characterized the aggregation properties of an experimentally determined bacterial proteome. The data reveal that the proteins that are effectively expressed in the bacterial cytosol and other compartments display aggregation properties that depend on the protein abundance, size, structure, function, relevance for cell fitness and specific localization in the cell. Overall, it appears that aggregation propensity acts as strong constraint during evolution, shaping different polypeptide properties. Accordingly, redundant natural mechanisms to avoid protein aggregation in biological contexts appear to exist. The data in the present study are consistent with many previous observations obtained through the analysis of theoretical proteomes using different computational strategies, which confirms the general validity of bioinformatic analyses to elucidate the mechanisms by which evolution tunes intrinsic protein aggregation properties. In turn, it is likely that the analysis of the aggregation properties of natural bacterial proteins would provide useful lessons to rationally manipulate and control the production of recombinant proteins in the bacterial cytosol.

Materials and Methods

Databases and parameters calculation

The amino acid sequences of bacterial proteins were obtained from Swiss-Prot Protein knowledgebase [78]. The protein subcellular location was obtained from PSORT database, version 2.0 [32,33].

The functions associated with the different sequences in the study were identified using the hierarchically structured functional catalogue (FunCat) [71] and the Protein Knowledgebase (UniProtKB) [72,73]. FunCat provides a set of functional categories, from 25 catalogued, for each classified protein. The biological processes associated with the different protein sets were assigned according to the ontology information in the TrEMBL database at the UniProtKB server. The essentiality of the bacterial proteins for the cellular fitness was derived from the data reported in [62,63].

The Database of Protein Disorder (DisProt) (release 4.9) has been used to identify disordered proteins or proteins containing extensive unstructured sequence stretches [66]. DisProt contains 47 *Escherichia coli* proteins experimentally shown to be intrinsically disordered; 20 of them are included in the analysed protein set.

The RegulonDB data base has been used to obtain the known *E. coli* operon structure set [79]. We only considered those operons encoding for at least 3 of the cytosolic proteins in the set.

The average hydropathy score (GRAVY) was calculated using the hydrophobicity values obtained from the Kyte-Doolittle scale [44]. The GRAVY was described as $(\sum^{\mathbf{n}}_{\mathbf{i=1}}\mathbf{H_i})/\mathbf{n}$ where **Hi** is the protein residue hydrophobicity at position **i** and **n** is the protein length.

The number of transmembrane regions was calculated employing TMHMM version 2.0 [70].

The Exponentially Modified Protein Abundance Index (emPAI) of each protein was obtained from the data reported in [29]. The cumulative distribution of the Na4vSS, NnHS and THSAr values associated with the 87 cytosolic polypeptides displaying the highest (MAP) and lowest emPAI (LAP) were plotted to analyse their aggregational properties. To analyse the overall correlation between cytosolic proteins abundance and their aggregation propensity we used the logarithm of emPAI LN(emPAI), because, as

Na4vSS, it displays in a lineal distribution. The LN(emPAI) comprise values between -2.5 and 23; however there were only 4 proteins between 14 and 23 values and they were discarded for further analysis. The remaining 872 proteins were divided in 45 grups at intervals of LN(emPAI) of 0.37 and the average value of each group calculated. In this way, the different length intervals have similar weights in the correlation, independently of the number of polypeptides present in each group.

The Codon Adaptation Index (CAI) values were obtained from [29]. The cytosolic polypeptides possess CAI values between 0.19 and 0.83. They were distributed in 20 intervals according to their CAI. Two of these intervals do not contain any protein or only one polypeptide and were discarded to avoid the dispersion of the data distribution. Subsequently the Na4vSS and CAI average of the 18 remainder groups were calculated.

The isoelectric points (pI) of the different polypeptides were calculated using the ProtParam tool of the ExPASy proteomics server of the Swiss Institute of Bioinformatics [78].

Composition of Hot Spots and flanquing stretches

Flanquing regions were defined as the 5 residues at the N- and C-sides of a given HS. The frequency of each natural amino acid inside the HSs and at their flanks was compared with their average frequency in natural proteins as deduced from Swiss-Prot [78]. The relative frequency of a given amino acid in HS (F_{rh}) was calculated as: **F_{rh}=(F_h/F_n)-1** where **F_f** is its frequency inside the HS and **F_n** its frequency in nature accordingly to Swiss-Prot data base [78]. Values above 1 or below 1 indicate higher or lower frequency, respectively. The same procedure was used to calculate the relative frequency of a given amino acid at the flanks.

Acknowledgements

We thank Monica Pardo for her assistance with the initial data analysis. We thank Prof. Aviles and Dr. Vendrell for lab facilities and continuous support. This work was supported by grants BIO2007-68046 from Ministerio de Ciencia y Innovación (Spain) and 2009-SGR 760 from AGAUR (Generalitat de

Catalunya). NSG was beneficiary of a FPI fellowship awarded by the Spanish Ministry.

Author contributions

Conceived and designed the experiments: SV. Performed the experiments: NSG. Analyzed the data: SV NSG. Wrote the paper: SV NSG.

References

1. Chiti F, Dobson CM (2006) Protein misfolding, functional amyloid, and human disease. Annu Rev Biochem 75: 333-366.

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

3. Fandrich M, Fletcher MA, Dobson CM (2001) Amyloid fibrils from muscle myoglobin. Nature 410: 165-166.

4. Guijarro JI, Sunde M, Jones JA, Campbell ID, Dobson CM (1998) Amyloid fibril formation by an SH3 domain. Proc Natl Acad Sci U S A 95: 4224-4228.

5. Carrio M, Gonzalez-Montalban N, Vera A, Villaverde A, Ventura S (2005) Amyloid-like properties of bacterial inclusion bodies. J Mol Biol 347: 1025-1037.

6. Wang L, Maji SK, Sawaya MR, Eisenberg D, Riek R (2008) Bacterial inclusion bodies contain amyloid-like structure. PLoS Biol 6: e195.

7. Wasmer C, Benkemoun L, Sabaté R, Steinmetz M, Coulary-Salin B, et al. (2009) Solid-State NMR reveals that E. coli inclusion bodies of HET-s(218-289) are amyloids. Angewandte Chemie Under review.

8. Morell M, Bravo R, Espargaro A, Sisquella X, Aviles FX, et al. (2008) Inclusion bodies: specificity in their aggregation process and amyloid-like structure. Biochim Biophys Acta 1783: 1815-1825.

9. Tartaglia GG, Vendruscolo M (2008) The Zyggregator method for predicting protein aggregation propensities. Chem Soc Rev 37: 1395-1401.

10. Conchillo-Sole O, de Groot NS, Aviles FX, Vendrell J, Daura X, et al. (2007) AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. BMC Bioinformatics 8: 65.

11. Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nat Biotechnol 22: 1302-1306.

12. Tartaglia GG, Cavalli A, Pellarin R, Caflisch A (2005) Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. Protein Sci 14: 2723-2734.

13. Zibaee S, Makin OS, Goedert M, Serpell LC (2007) A simple algorithm locates beta-strands in the amyloid fibril core of alpha-synuclein, Abeta, and tau using the amino acid sequence alone. Protein Sci 16: 906-918.

14. Bryan AW, Jr., Menke M, Cowen LJ, Lindquist SL, Berger B (2009) BETASCAN: probable beta-amyloids identified by pairwise probabilistic analysis. PLoS Comput Biol 5: e1000333.

15. Rojas Quijano FA, Morrow D, Wise BM, Brancia FL, Goux WJ (2006) Prediction of nucleating sequences from amyloidogenic propensities of tau-related peptides. Biochemistry 45: 4638-4652.

16. Trovato A, Chiti F, Maritan A, Seno F (2006) Insight into the structure of amyloid fibrils from the analysis of globular proteins. PLoS Comput Biol 2: e170.

17. Saiki M, Konakahara T, Morii H (2006) Interaction-based evaluation of the propensity for amyloid

formation with cross-beta structure. Biochem Biophys Res Commun 343: 1262-1271.

18. Thompson MJ, Sievers SA, Karanicolas J, Ivanova MI, Baker D, et al. (2006) The 3D profile method for identifying fibril-forming segments of proteins. Proc Natl Acad Sci U S A 103: 4074-4078.

19. Galzitskaya OV, Garbuzynskiy SO, Lobanov MY (2006) Prediction of amyloidogenic and disordered regions in protein chains. PLoS Comput Biol 2: e177.

20. Yoon S, Welsh WJ (2004) Detecting hidden sequence propensity for amyloid fibril formation. Protein Sci 13: 2149-2160.

21. Monsellier E, Ramazzotti M, Taddei N, Chiti F (2008) Aggregation propensity of the human proteome. PLoS Comput Biol 4: e1000199.

22. Linding R, Schymkowitz J, Rousseau F, Diella F, Serrano L (2004) A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins. J Mol Biol 342: 345-353.

23. Tartaglia GG, Caflisch A (2007) Computational analysis of the S. cerevisiae proteome reveals the function and cellular localization of the least and most amyloidogenic proteins. Proteins 68: 273-278.

24. Tartaglia GG, Pellarin R, Cavalli A, Caflisch A (2005) Organism complexity anti-correlates with proteomic beta-aggregation propensity. Protein Sci 14: 2735-2740.

25. de Groot NS, Aviles FX, Vendrell J, Ventura S (2006) Mutagenesis of the central hydrophobic cluster in Abeta42 Alzheimer's peptide. Side-chain properties correlate with aggregation propensities. Febs J 273: 658-668.

26. Ventura S, Zurdo J, Narayanan S, Parreno M, Mangues R, et al. (2004) Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case. Proc Natl Acad Sci U S A 101: 7258-7263.

27. Rousseau F, Serrano L, Schymkowitz JW (2006) How evolutionary pressure against protein aggregation shaped chaperone specificity. J Mol Biol 355: 1037-1047.

28. Sanchez de Groot N, Pallares I, Aviles FX, Vendrell J, Ventura S (2005) Prediction of "hot spots" of aggregation in disease-linked polypeptides. BMC Struct Biol 5: 18.

29. Ishihama Y, Schmidt T, Rappsilber J, Mann M, Hartl FU, et al. (2008) Protein abundance profiling of the Escherichia coli cytosol. BMC Genomics 9: 102.

30. de Groot NS, Ventura S (2006) Protein activity in bacterial inclusion bodies correlates with predicted aggregation rates. J Biotechnol 125: 110-113.

31. de Groot NS, Ventura S (2006) Effect of temperature on protein quality in bacterial inclusion bodies. FEBS Lett 580: 6471-6476.

32. Gardy JL, Laird MR, Chen F, Rey S, Walsh CJ, et al. (2005) PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics 21: 617-623 URL: http://db.psort.org/.

33. Rey S, Acab M, Gardy JL, Laird MR, deFays K, et al. (2005) PSORTdb: a protein subcellular localization database for bacteria. Nucleic Acids Res 33: D164-168.

34. Selinger DW, Cheung KJ, Mei R, Johansson EM, Richmond CS, et al. (2000) RNA expression analysis using a 30 base pair resolution Escherichia coli genome array. Nat Biotechnol 18: 1262-1268.

35. Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, et al. (2005) Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. Mol Cell Proteomics 4: 1265-1272.

36. Wilks JC, Slonczewski JL (2007) pH of the cytoplasm and periplasm of Escherichia coli: rapid measurement by green fluorescent protein fluorimetry. J Bacteriol 189: 5601-5607.

37. Chiti F, Calamai M, Taddei N, Stefani M, Ramponi G, et al. (2002) Studies of the aggregation of mutant proteins in vitro provide insights into the genetics of amyloid diseases. Proc Natl Acad Sci U S A 99 Suppl 4: 16419-16426.

38. Vetri V, Librizzi F, Leone M, Militello V (2007) Thermal aggregation of bovine serum albumin at different pH: comparison with human serum albumin. Eur Biophys J 36: 717-725.

39. Militello V, Casarino C, Emanuele A, Giostra A, Pullara F, et al. (2004) Aggregation kinetics of

bovine serum albumin studied by FTIR spectroscopy and light scattering. Biophys Chem 107: 175-187.

40. Shirota M, Ishida T, Kinoshita K (2008) Effects of surface-to-volume ratio of proteins on hydrophilic residues: decrease in occurrence and increase in buried fraction. Protein Sci 17: 1596-1602.

41. Lawrence MS, Phillips KJ, Liu DR (2007) Supercharging proteins can impart unusual resilience. J Am Chem Soc 129: 10110-10112.

42. Vendruscolo M, Dobson CM (2007) Chemical biology: More charges against aggregation. Nature 449: 555.

43. de Groot NS, Parella T, Aviles FX, Vendrell J, Ventura S (2007) Ile-phe dipeptide self-assembly: clues to amyloid formation. Biophys J 92: 1732-1741.

44. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157: 105-132.

45. Seo MJ, Jeong KJ, Leysath CE, Ellington AD, Iverson BL, et al. (2009) Engineering antibody fragments to fold in the absence of disulfide bonds. Protein Sci 18: 259-267.

46. Sharp PM, Li WH (1987) The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15: 1281-1295.

47. Jansen R, Bussemaker HJ, Gerstein M (2003) Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. Nucleic Acids Res 31: 2242-2251.

48. Tartaglia GG, Pechmann S, Dobson CM, Vendruscolo M (2009) A relationship between mRNA expression levels and protein solubility in E. coli. J Mol Biol 388: 381-389.

49. Shen M-y, Davis, F. P., Sali, A. (2005) The optimal size of a globular protein domain: A simple sphere-packing model. Chemical Physics Letters 405: 224-228.

50. Teller DC (1976) Accessible area, packing volumes and interaction surfaces of globular proteins. Nature 260: 729-731.

51. Sandelin E (2004) On hydrophobicity and conformational specificity in proteins. Biophys J 86: 23-30.

52. Irback A, Sandelin E (2000) On hydrophobicity correlations in protein chains. Biophys J 79: 2252-2258.

53. Kajander T, Kahn PC, Passila SH, Cohen DC, Lehtio L, et al. (2000) Buried charged surface in proteins. Structure 8: 1203-1214.

54. Bolon DN, Mayo SL (2001) Polar residues in the protein core of Escherichia coli thioredoxin are important for fold specificity. Biochemistry 40: 10047-10053.

55. Kiraga J, Mackiewicz P, Mackiewicz D, Kowalczuk M, Biecek P, et al. (2007) The relationships between the isoelectric point and: length of proteins, taxonomy and ecology of organisms. BMC Genomics 8: 163.

56. Ivankov DN, Garbuzynskiy SO, Alm E, Plaxco KW, Baker D, et al. (2003) Contact order revisited: influence of protein size on the folding rate. Protein Sci 12: 2057-2062.

57. Ellis RJ (2000) Chaperone substrates inside the cell. Trends Biochem Sci 25: 210-212.

58. Thulasiraman V, Yang CF, Frydman J (1999) In vivo newly translated polypeptides are sequestered in a protected folding environment. EMBO J 18: 85-95.

59. Srikakulam R, Winkelmann DA (1999) Myosin II folding is mediated by a molecular chaperonin. J Biol Chem 274: 27265-27273.

60. Monsellier E, Chiti F (2007) Prevention of amyloid-like aggregation as a driving force of protein evolution. EMBO Rep 8: 737-742.

61. Reumers J, Maurer-Stroh S, Schymkowitz J, Rousseau F (2009) Protein sequences encode safeguards against aggregation. Hum Mutat 30: 431-437.

62. Gerdes SY, Scholle MD, Campbell JW, Balazsi G, Ravasz E, et al. (2003) Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. J Bacteriol 185: 5673-5684.

63. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, et al. (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol 2: 2006 0008.

64. Chen Y, Dokholyan NV (2008) Natural selection against protein aggregation on self-interacting and

essential proteins in yeast, fly, and worm. Mol Biol Evol 25: 1530-1533.

65. Chen JW, Romero P, Uversky VN, Dunker AK (2006) Conservation of intrinsic disorder in protein domains and families: II. functions of conserved disorder. J Proteome Res 5: 888-898.

66. Vucetic S, Obradovic Z, Vacic V, Radivojac P, Peng K, et al. (2005) DisProt: a database of protein disorder. Bioinformatics 21: 137-140.

67. Santoni V, Molloy M, Rabilloud T (2000) Membrane proteins and proteomics: un amour impossible? Electrophoresis 21: 1054-1070.

68. Dougan DA, Mogk A, Bukau B (2002) Protein folding and degradation in bacteria: to degrade or not to degrade? That is the question. Cell Mol Life Sci 59: 1607-1616.

69. Liu Y, Fu X, Shen J, Zhang H, Hong W, et al. (2004) Periplasmic proteins of Escherichia coli are highly resistant to aggregation: reappraisal for roles of molecular chaperones in periplasm. Biochem Biophys Res Commun 316: 795-801.

70. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305: 567-580 URL: http://www.cbs.dtu.dk/services/TMHMM/

71. Ruepp A, Zollner A, Maier D, Albermann K, Hani J, et al. (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res 32: 5539-5545.

72. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, et al. (2004) UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 32: D115-119.

73. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, et al. (2005) The Universal Protein Resource (UniProt). Nucleic Acids Res 33: D154-159.

74. Alix E, Blanc-Potard AB (2009) Hydrophobic peptides: novel regulators within bacterial membrane. Mol Microbiol 72: 5-11.

75. Cowan SW, Schirmer T, Rummel G, Steiert M, Ghosh R, et al. (1992) Crystal structures explain functional properties of two E. coli porins. Nature 358: 727-733.

76. Schirmer T (1998) General and specific porins from bacterial outer membranes. J Struct Biol 121: 101-109.

77. Knowles TJ, Scott-Tucker A, Overduin M, Henderson IR (2009) Membrane protein architects: the role of the BAM complex in outer membrane protein assembly. Nat Rev Microbiol 7: 206-214.

78. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31: 365-370 URL: http://www.expasy.ch/sprot/.

79. Huerta AM, Salgado H, Thieffry D, Collado-Vides J (1998) RegulonDB: a database on transcriptional regulation in Escherichia coli. Nucleic Acids Res 26: 55-59.

Figure legends

Figure 1. Example of AGGRESCAN output. The red line represents the aggregation profile of a putative protein with 35 amino acids. The blue line indicates the Hot Spot (HS) threshold, according to the individual aggregation propensity of the 20 natural amino acids and their frequency in natural proteins [28]. The green line corresponds to the average aggregation propensity of the putative protein. The HS areas over the threshold are filled in red (A and B). av4 is the aggregation propensity average over a sliding window of 5 to 11 residues [10]. The aggregation propensity of each amino acid results from the depositional analysis of a set of amyloid polypeptides in the *E. coli* cytoplasm [25,28].

Figure 2. Relationship between the cytosolic proteins abundance and the AGGRESCAN aggregation parameters. Cumulative distributions of the NnHS (A), THSAr (B) and Na4vSS (C) parameters in the 10% most abundant cytosolic proteins (MAP, black) and the 10% least abundant ones (LAP, grey). D) Correlation between protein abundance, measured as LN(emPAI), and protein aggregation propensity, measured as Na4vSS, in the complete cytosolic protein set. The analysed proteins were divided in 45 grups according to their LN(emPAI) values. Each point in the graphic represents the average value of the corresponding group.

Figure 3. Relationship between the cytosolic proteins abundance and their intrinsic properties. A) Amino acid abundance in MAP (pale grey) and LAP (dark grey) sequences relative to the expected frequencies in natural proteins as deduced from Swiss-Prot [78]. B) Comparison between the proteins pI and Na4vSS values. C) Correlation between proteins hydropathicity (GRAVY) and Na4vSS values.

Figure 4. Comparison between cytosolic proteins theoretical expression levels and their aggregation parameters. A) Cumulative distributions of Na4vSS values in the 10% cytosolic proteins with the highest (black) and lowest (grey) CAI values. B) Correlation between the CAI and the Na4vSS values. Each point represents the average value over all the sequences having a CAI value comprised in an interval of 0.03.

Figure 5. Dependence of proteins length on their aggregation properties and chaperone binding affinity. A) Dot plot distribution represents the relationship between the molecular weight and Na4vSS. Columns show the size distribution of polypeptides that bind to GroEL

(grey) or DnaK (white) in *E. coli.* according to the data in [57]. B) Relationship between the molecular weight and the NnHS. Each point corresponds to the average value over all the sequences having a length comprised in an interval of 1.9 kDa.

Figure 6. Amino acid composition of cytosolic proteins HSs and their flanks. A) Amino acid frequencies relative to their average frequency in natural proteins as deduced from Swiss-Prot [78]. A relative frequency of 0 for a given residue at a given position means that the residue occupies that position with a frequency identical to that in natural proteins. Residues enrichment in the HSs (B) and at the flanks (C) relative to their frequency in natural proteins. Values above or below 1.0 point denote increases or decreases in frequency, respectively.

Figure 7. Proteins encoded by the same operon display related aggregation propensities. Standard deviation of Na4vSS values in the 25 analysed operons. The standard deviation in the complete cytosolic set is 7.72 (dashed line). Low standard deviation within an operon indicates that the aggregation propensity of its proteins is similar.

Figure 8. Disordered sequence stretches display reduced protein aggregation. Cumulative distributions of NnHS and Na4vSS values in ribosomal proteins (A and B), intrinsically unstructured proteins (C and D) and disordered fragments in cytosolic proteins (E and F) are compared with the distribution in the complete cytosolic set (grey).

Figure 9. Relationship between subcellular localisation and protein aggregation propensity. Cumulative distribution of NnHS (A), THSAr (B) and Na4vSS (C) of proteins located in the cytoplasm (C, red), outer membrane (OM, dark green), periplasm (P, blue). D) Dot distribution of the Na4vSS values of the proteins in the previous four protein sets as well as those located in the inner membrane (IM, pale green); the vertical lines correspond to the Na4vSS mean in each protein set. Cumulative distribution of NnHS (E) and Na4vSS (F) in cytosolic and inner membrane proteins.

Figure 10. The inner membrane contains proteins with different number of transmembrane segments and associated aggregation propensities. Diagram of the inner membrane protein set showing the Na4vSS value and the number of transmembrane segments.

Figure 11. Inner membrane proteins with differential aggregation propensities are involved in different biological functions. Percentage of inner membrane proteins associated

with the biological functions described in FunCat (A) and UniProtKB (B). The inner membrane proteins were divided in two groups according to their Na4vSS value: Na4vSS < 6 (42 proteins; pale grey) or Na4vSS ≥ 6 (43 proteins; dark grey).

Table 1. Different operons regulate proteins with different aggregation propensity and biological function.

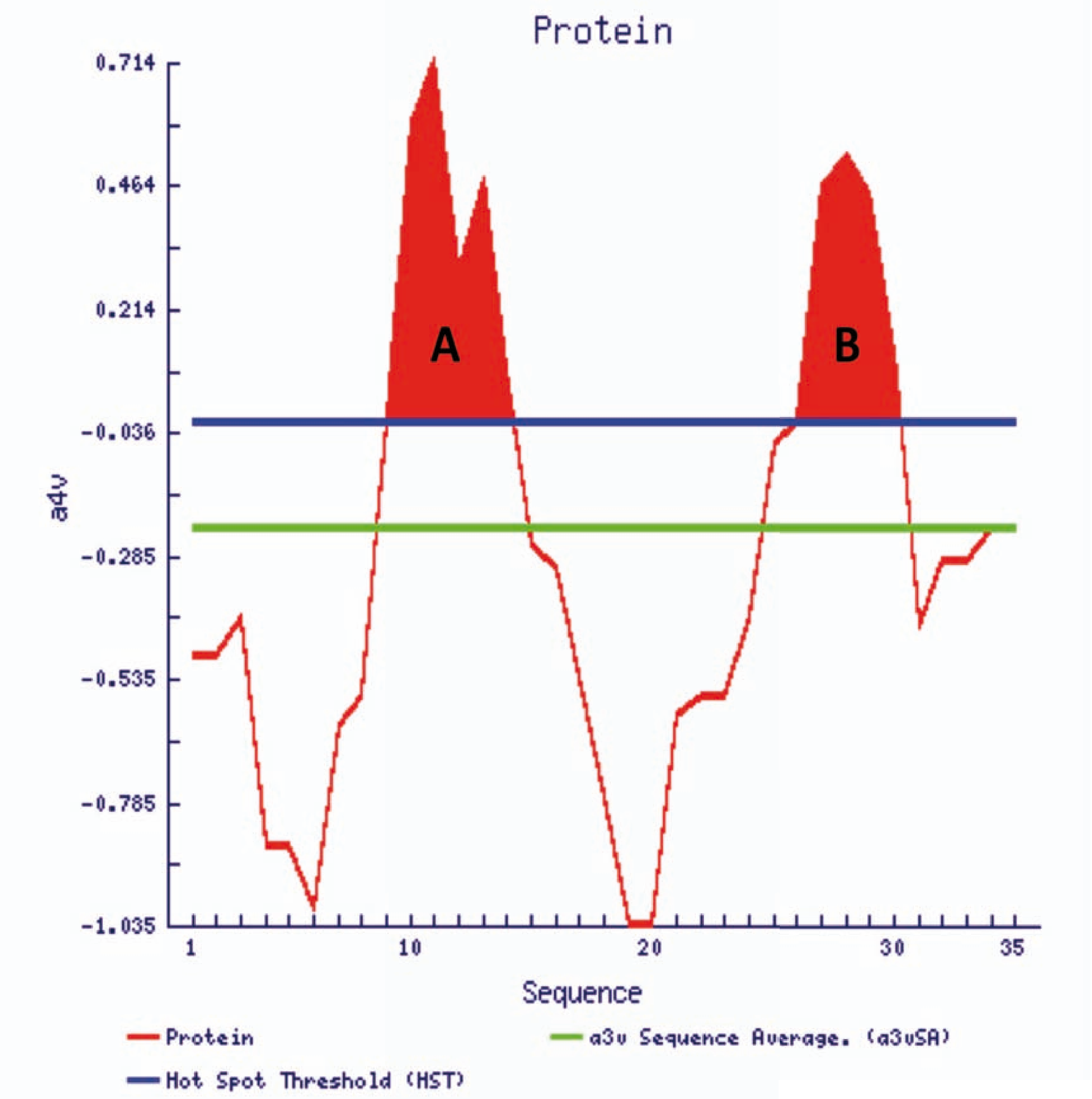
LA operons name[a]	Na4vSS	nº proteins	Ribosomal	Essential	Non-essential	Unknow
yjeFE-amiB-mutL-miaA-hfq-hflXKC	-15.63	3	0	1	2	0
hscBA-fdx	-14.33	3	0	2	0	1
rpsMKD-rpoA-rplQ	-14.32	5	4	3	0	2
cmk-rpsA-himD	-13.20	3	1	0	0	2
rpsF-priB-rpsR-rplI	-12.93	3	3	2	0	1
pheST-himA	-12.50	3	0	0	0	3
rpsLG-fusA-tufA	-11.70	3	2	2	0	1
rpsJ-rplCDWB-rpsS-rplV-rpsC-rplP-rpmC-rpsQ	-11.47	11	11	4	0	7
thrS-infC-rpmI-rplT	-11.25	4	2	0	0	4
metY-yhbC-nusA-infB-rbfA-truB-rpsO-pnp	-11.17	7	1	4	0	3
iscRSUA	-9.78	4	0	2	1	1
rpsP-rimM-trmD-rplS	-8.60	4	2	3	0	1
rplNXE-rpsNH-rplFR-rpsE-rpmD-rplO-prlA-rpmJ	-8.52	9	9	3	0	6
aroKB-damX-dam-rpe-gph-trpS	-7.60	3	0	2	0	1
galETKM	-7.47	3	0	0	2	1
	Total	**68**	**35**	**28**	**5**	**34**
		%	**51.47**	**41.18**	**7.35**	**50.00**

HA operons name[b]	Na4vSS	nº proteins	Ribosomal	Essential	Non-essential	Unknow
ribF-ileS-lspA-slpA-lytB	-5.97	3	0	1	1	1
rplJL-rpoBC	-5.93	4	2	0	0	4
nuoABCEFGHIJKLMN	-5.87	3	0	0	1	2
sdhCDAB-b0725-sucABCD	-5.74	5	0	2	2	1
leuLABCD	-5.55	4	0	0	0	4
entCEBA-ybdB	-5.54	5	0	0	4	1
minCDE	-4.50	3	0	2	0	1
fabHDG-acpP-fabF	-4.38	4	0	4	0	0
gcvTHP	-4.13	3	0	0	0	3
dhaKLM	-4.03	3	0	1	0	2
ptsHI-crr	-3.33	3	0	0	1	2
deoCABD	-3.23	4	0	0	1	3
thiCEFGH	-2.53	4	0	0	1	3
hisGDCBHAFI	-1.87	3	0	0	0	3
mraZW-ftsLI-murEF-mraY-murD-ftsW-murGC-ddlB-ftsQAZ	-1.68	4	0	3	0	1
rfbBDACX	-0.86	5	0	0	3	2
gatYZABCDR_2	5.90	4	0	1	1	2
	Total	**64**	**2**	**14**	**15**	**35**
		%	**3.13**	**21.88**	**23.44**	**54.69**

a. Operons regulating proteins with aggregation propensity lower (LA) than the mean aggregation propensity of the complete operon protein set (-6.4 Na4vSS).

b. Operons regulating proteins with aggregation propensity higher (HA) than the mean aggregation propensity of the complete operon protein set (-6.4 Na4vSS).

Figure 1



(2HS/35 residues)·100 = **5.71 NnHS**

(A+B)/35 residues = **0.11 THSAr**

($\bar{X}$a4v)·100 = **-26.8 Na4vSS**

Figure 2

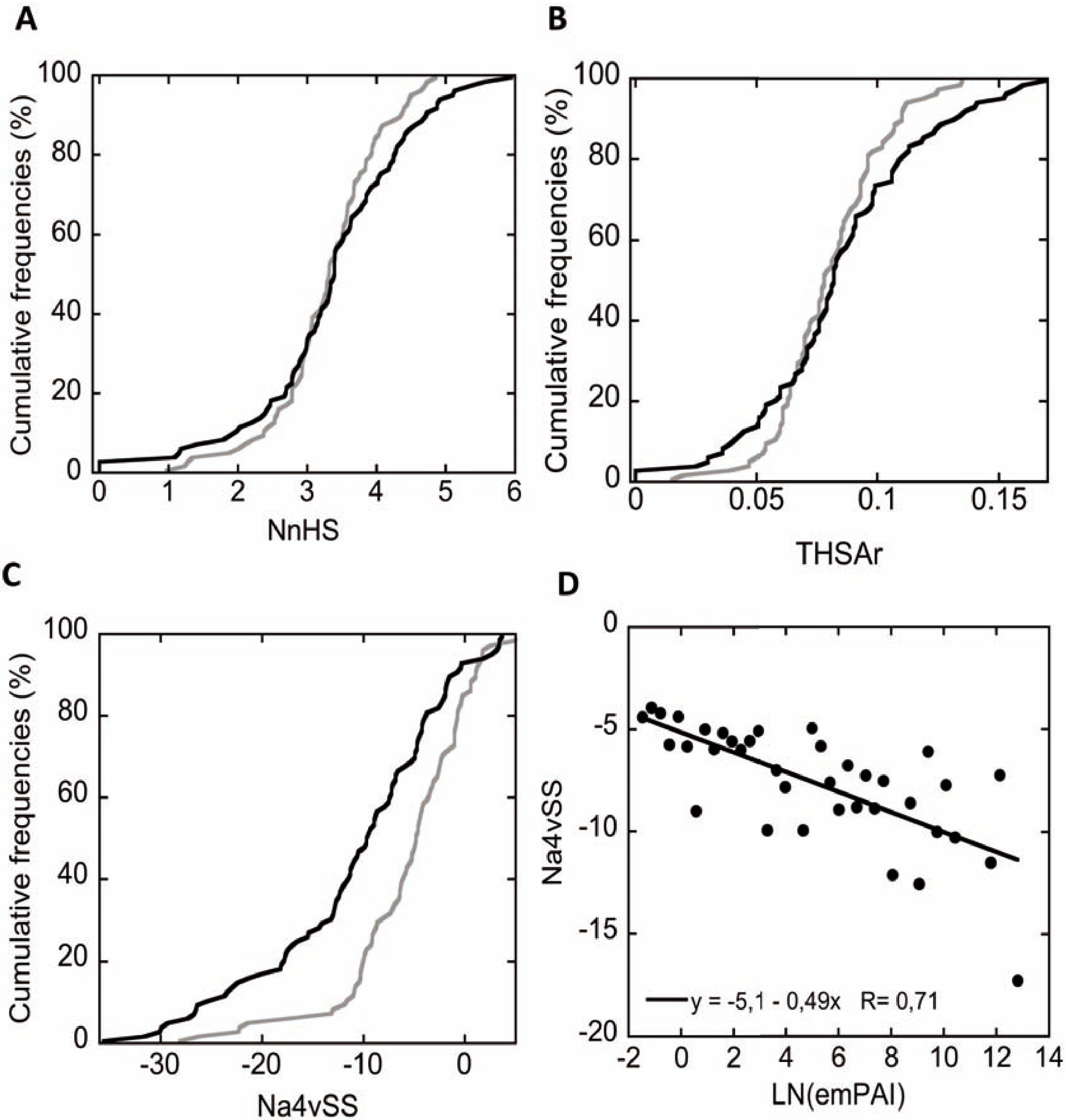

A
Cumulative frequencies (%)
100
80
60
40
20
0
0
1
2
3
4
5
6
NnHS
B
Cumulative frequencies (%)
100
80
60
40
20
0
0
0.05
0.1
0.15
THSAr
C
Cumulative frequencies (%)
100
80
60
40
20
0
-30
-20
-10
0
Na4vSS
D
0
-5
-10
-15
-20
Na4vSS
-2
0
2
4
6
8
10
12
14
LN(emPAI)
y = -5,1 - 0,49x R= 0,71

Figure 3

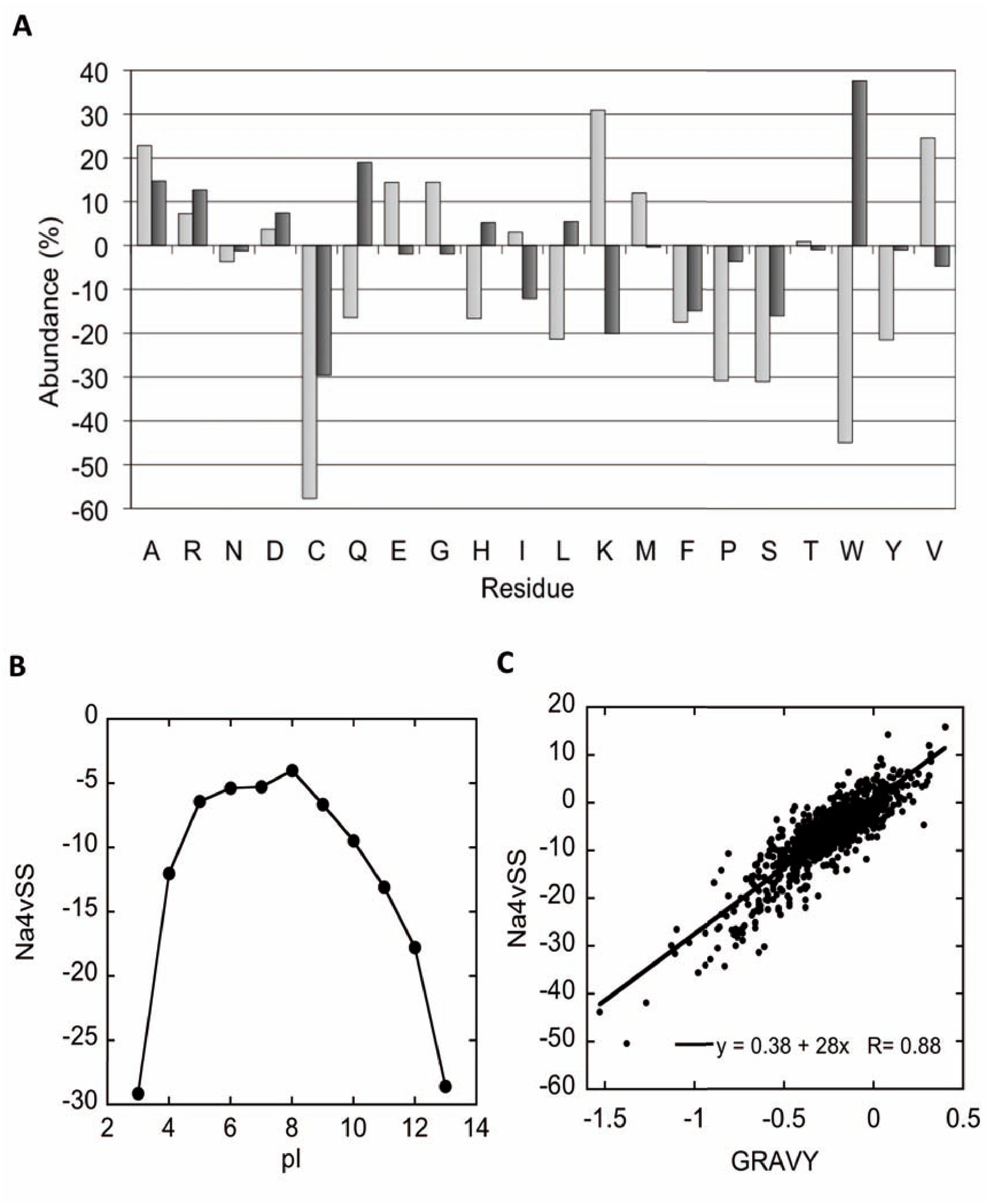

A
40
30
20
10
0
-10
-20
-30
-40
-50
-60
Abundance (%)
A R N D C Q E G H I L K M F P S T W Y V
Residue
B
0
-5
-10
-15
-20
-25
-30
Na4vSS
2 4 6 8 10 12 14
pI
C
20
10
0
-10
-20
-30
-40
-50
-60
Na4vSS
-1.5 -1 -0.5 0 0.5
GRAVY
y = 0.38 + 28x R= 0.88

Figure 4

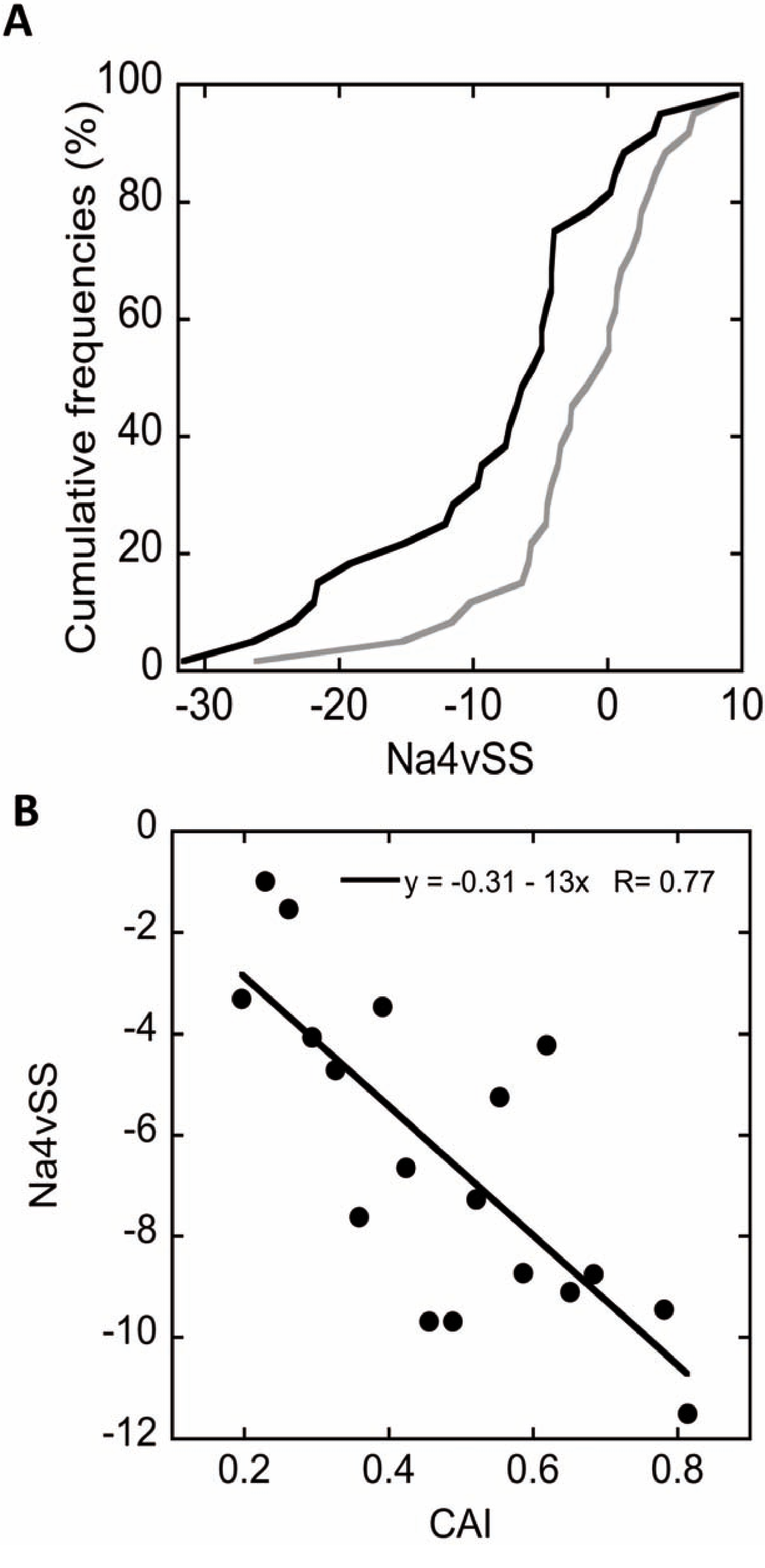

A
Cumulative frequencies (%)
100
80
60
40
20
0
-30
-20
-10
0
10
Na4vSS
B
0
-2
-4
-6
-8
-10
-12
Na4vSS
y = -0.31 - 13x R= 0.77
0.2
0.4
0.6
0.8
CAI

Figure 5

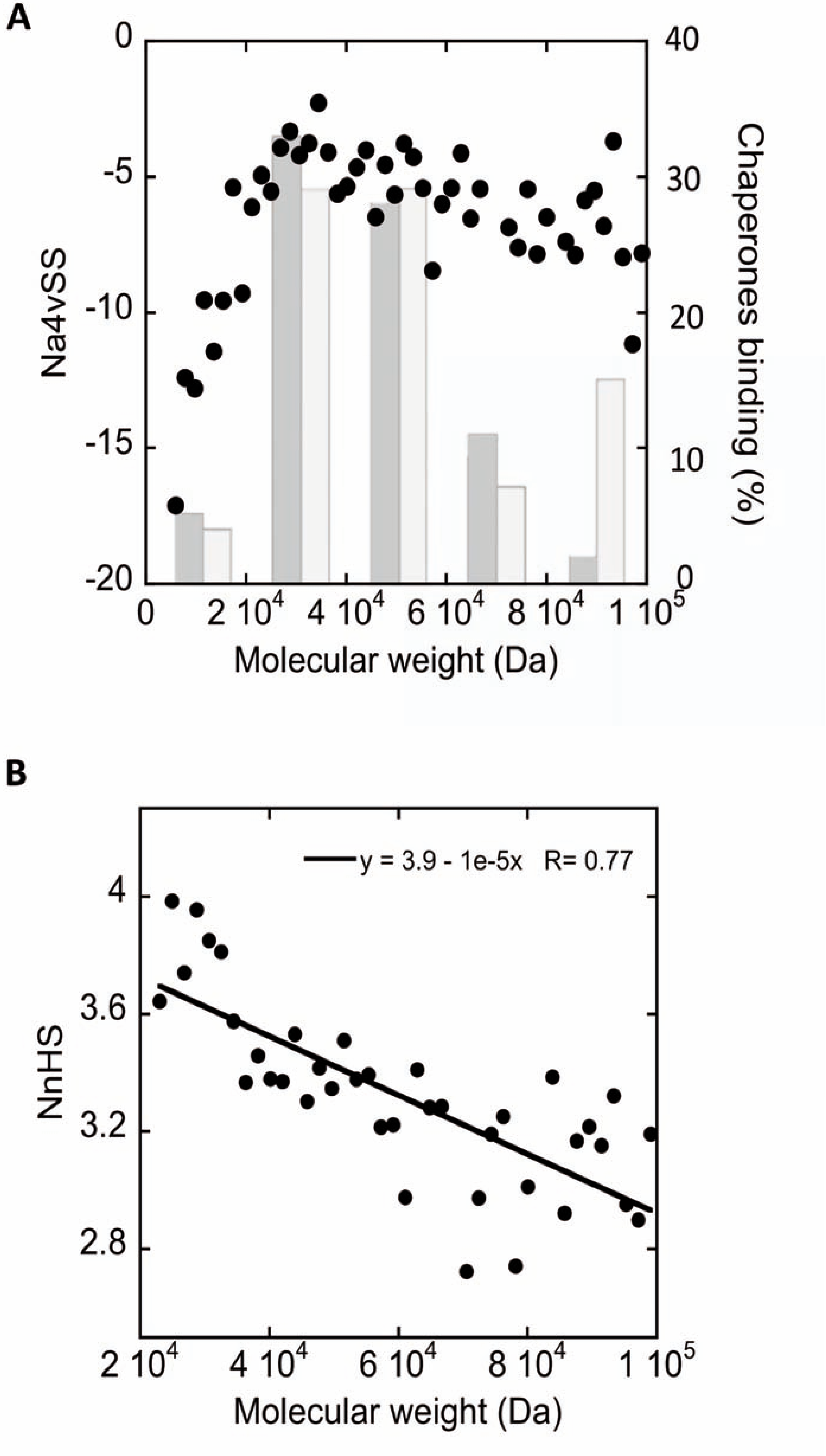

A
0
-5
-10
-15
-20
Na4vSS
40
30
20
10
0
Chaperones binding (%)
0
2 10^4
4 10^4
6 10^4
8 10^4
1 10^5
Molecular weight (Da)
B
y = 3.9 - 1e-5x   R= 0.77
4
3.6
3.2
2.8
NnHS
2 10^4
4 10^4
6 10^4
8 10^4
1 10^5
Molecular weight (Da)

Figure 6

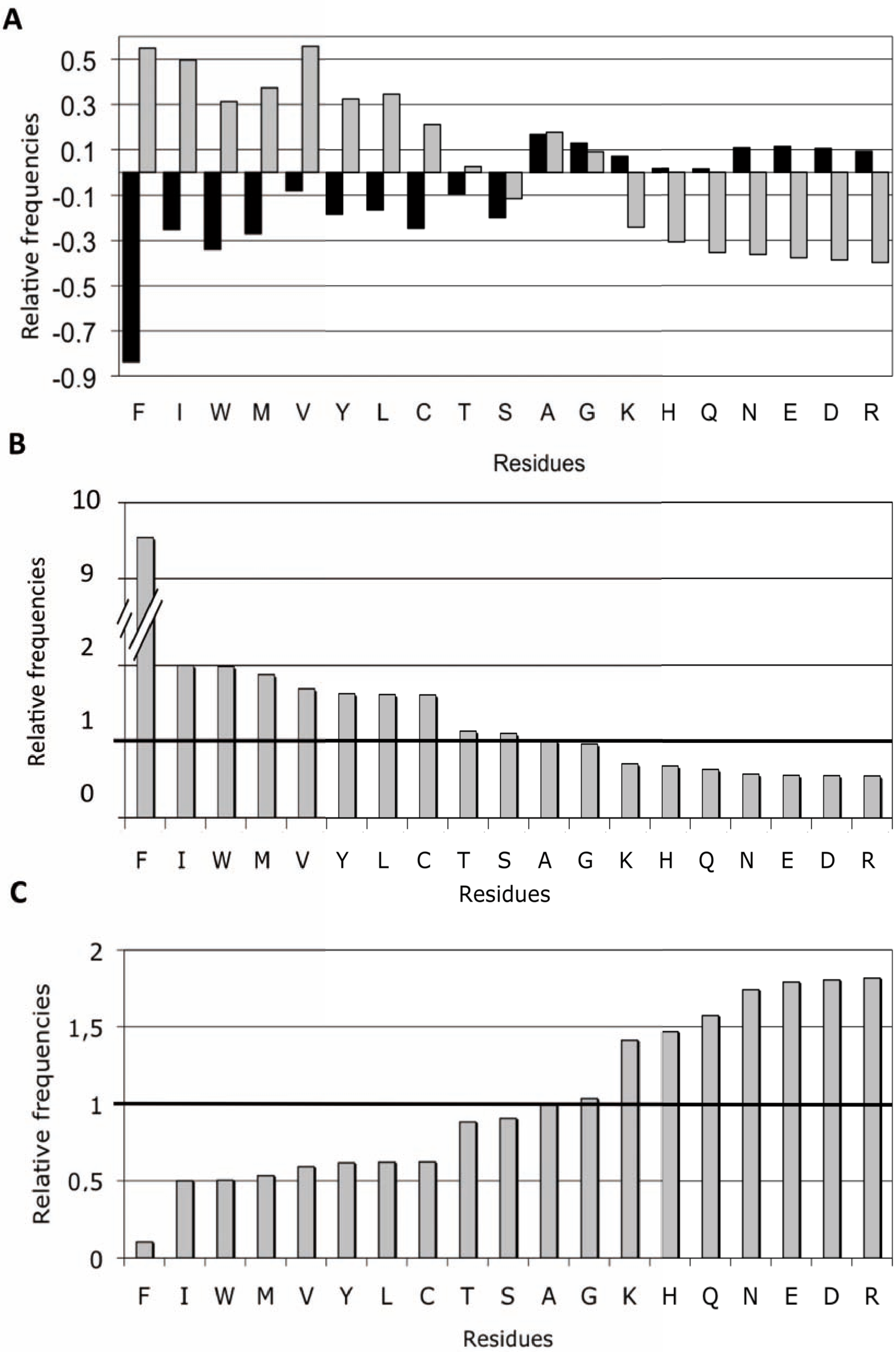

A
Relative frequencies
0.5
0.3
0.1
-0.1
-0.3
-0.5
-0.7
-0.9
F I W M V Y L C T S A G K H Q N E D R
Residues
B
Relative frequencies
10
9
2
1
0
F I W M V Y L C T S A G K H Q N E D R
Residues
C
Relative frequencies
2
1,5
1
0,5
0
F I W M V Y L C T S A G K H Q N E D R
Residues

Figure 7

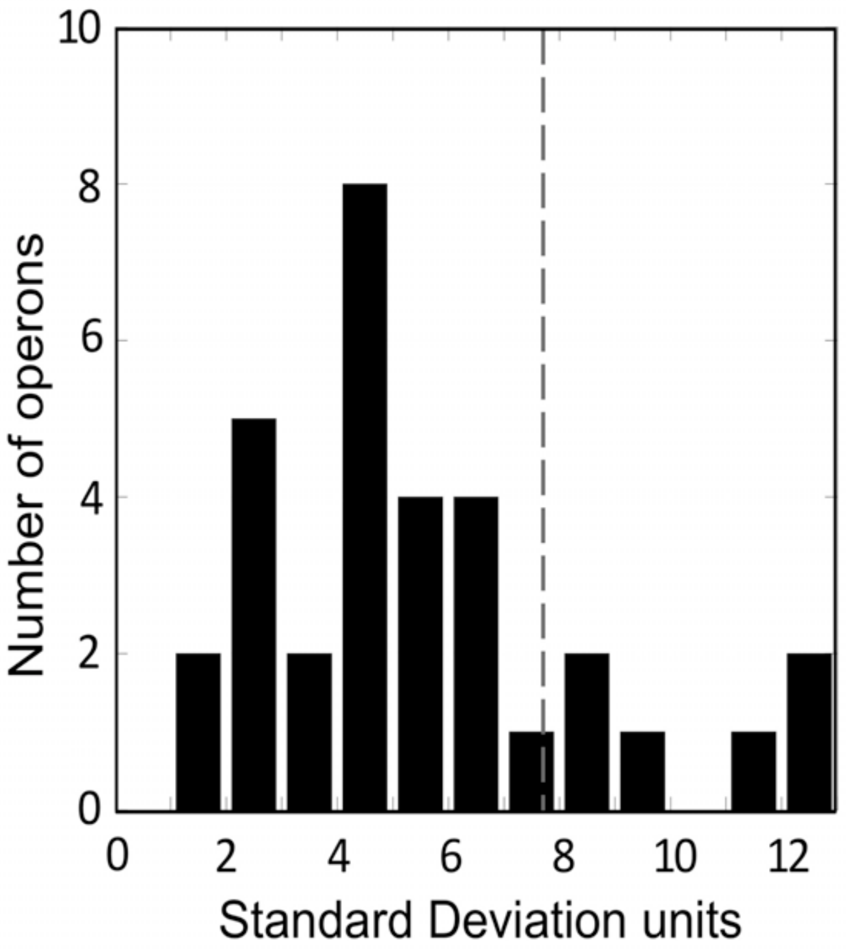

10
8
6
4
2
0
Number of operons
0
2
4
6
8
10
12
Standard Deviation units

Figure 8

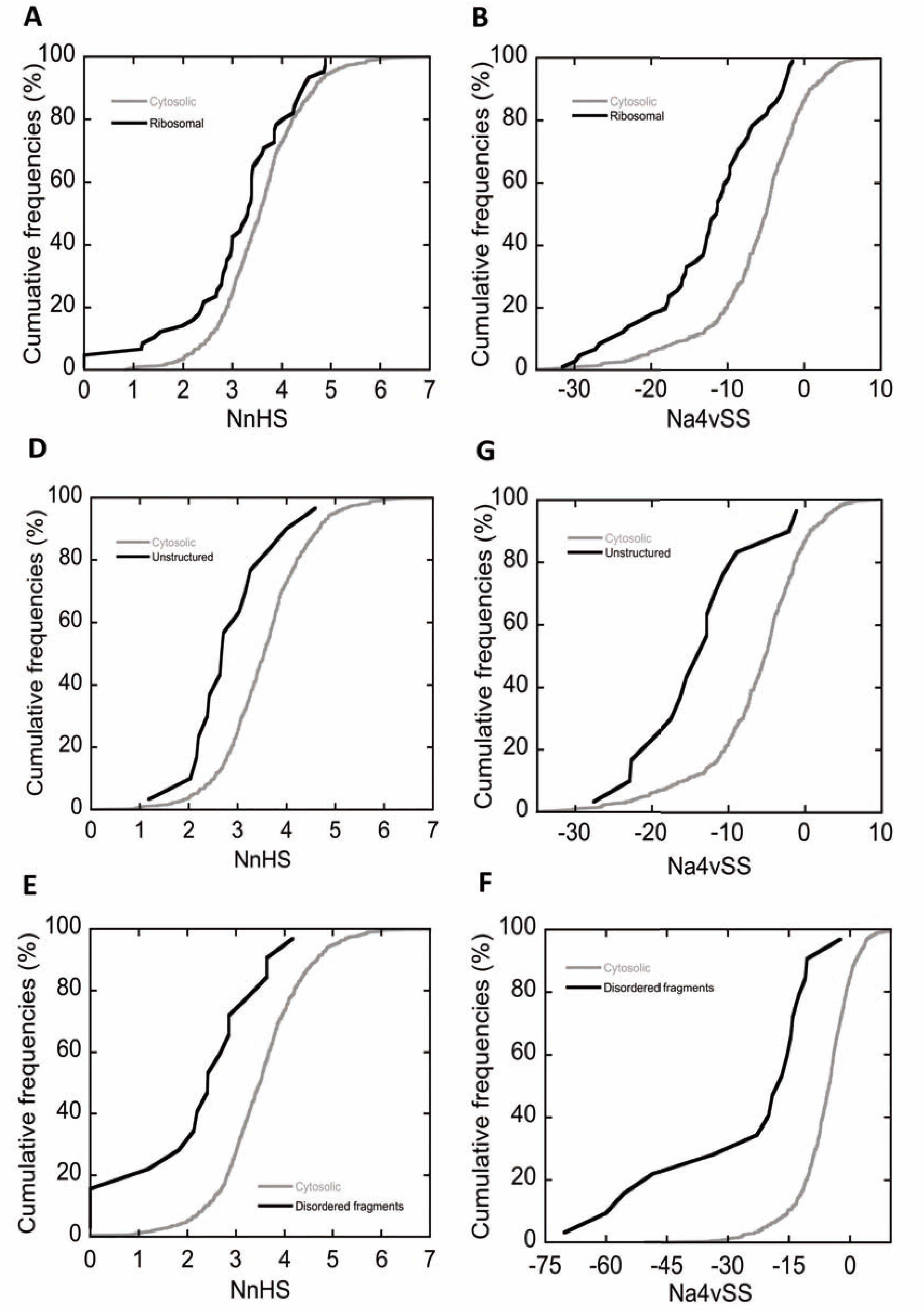

A
Cumulative frequencies (%)
100
80
60
40
20
0
Cytosolic
Ribosomal
0
1
2
3
4
5
6
7
NnHS
B
Cumulative frequencies (%)
100
80
60
40
20
0
Cytosolic
Ribosomal
-30
-20
-10
0
10
Na4vSS
D
Cumulative frequencies (%)
100
80
60
40
20
0
Cytosolic
Unstructured
0
1
2
3
4
5
6
7
NnHS
G
Cumulative frequencies (%)
100
80
60
40
20
0
Cytosolic
Unstructured
-30
-20
-10
0
10
Na4vSS
E
Cumulative frequencies (%)
100
80
60
40
20
0
Cytosolic
Disordered fragments
0
1
2
3
4
5
6
7
NnHS
F
Cumulative frequencies (%)
100
80
60
40
20
0
Cytosolic
Disordered fragments
-75
-60
-45
-30
-15
0
Na4vSS

Figure 9

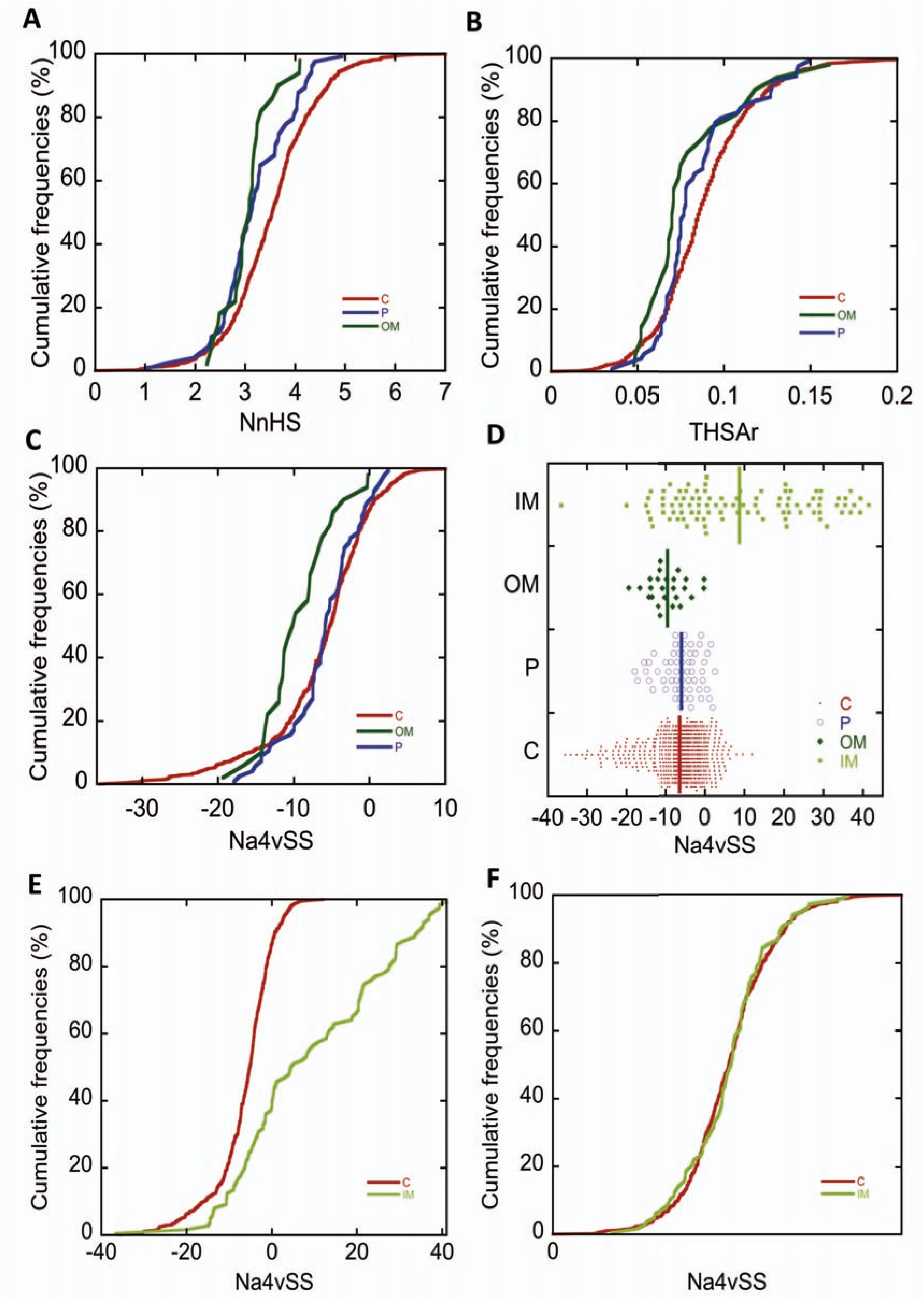

A
Cumulative frequencies (%)
0
20
40
60
80
100
0
1
2
3
4
5
6
7
NnHS
C
P
OM
B
Cumulative frequencies (%)
0
20
40
60
80
100
0
0.05
0.1
0.15
0.2
THSAr
C
OM
P
C
Cumulative frequencies (%)
0
20
40
60
80
100
-30
-20
-10
0
10
Na4vSS
C
OM
P
D
IM
OM
P
C
-40
-30
-20
-10
0
10
20
30
40
Na4vSS
C
P
OM
IM
E
Cumulative frequencies (%)
0
20
40
60
80
100
-40
-20
0
20
40
Na4vSS
C
IM
F
Cumulative frequencies (%)
0
20
40
60
80
100
0
Na4vSS
C
IM

Figure 10

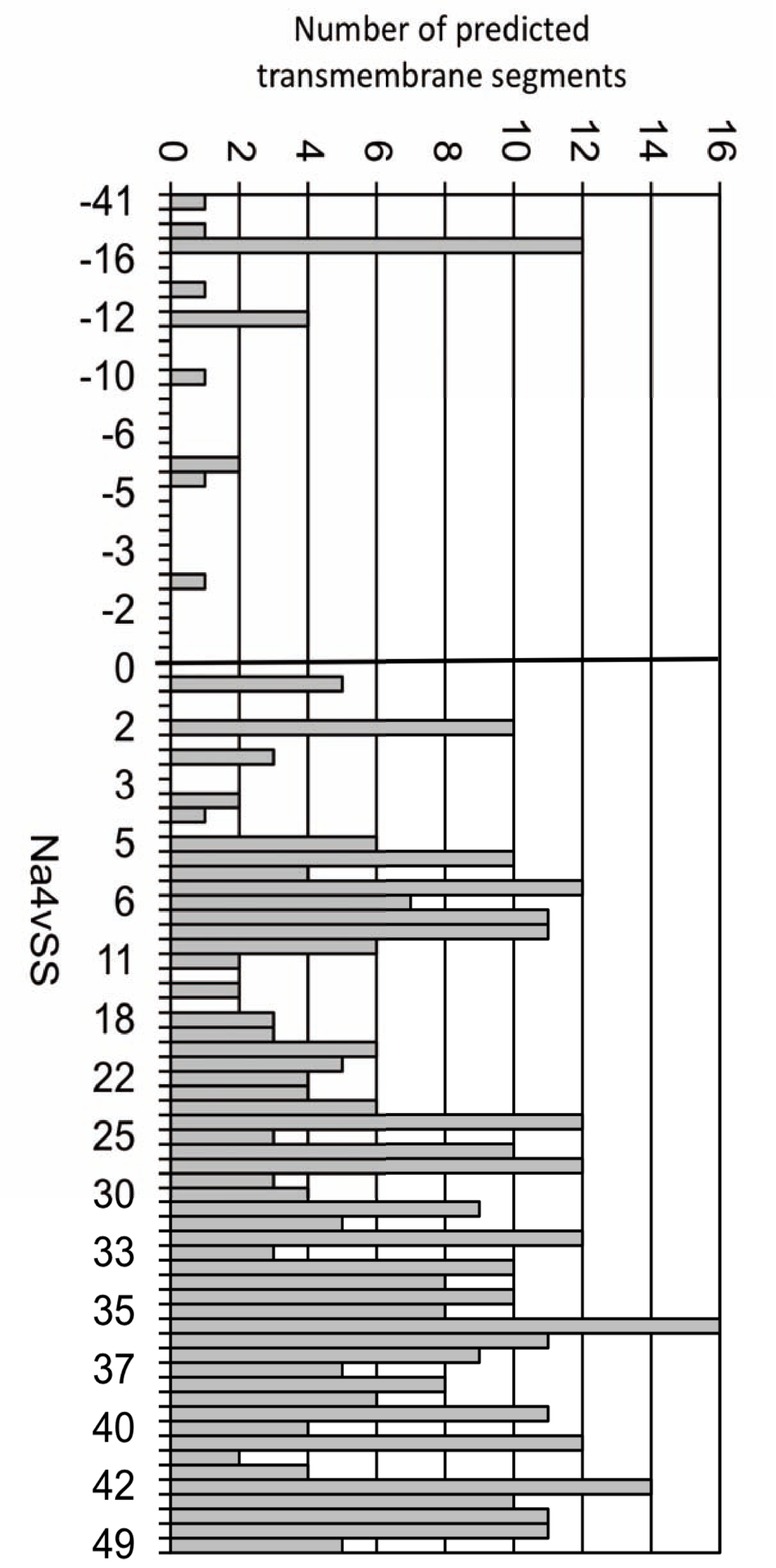
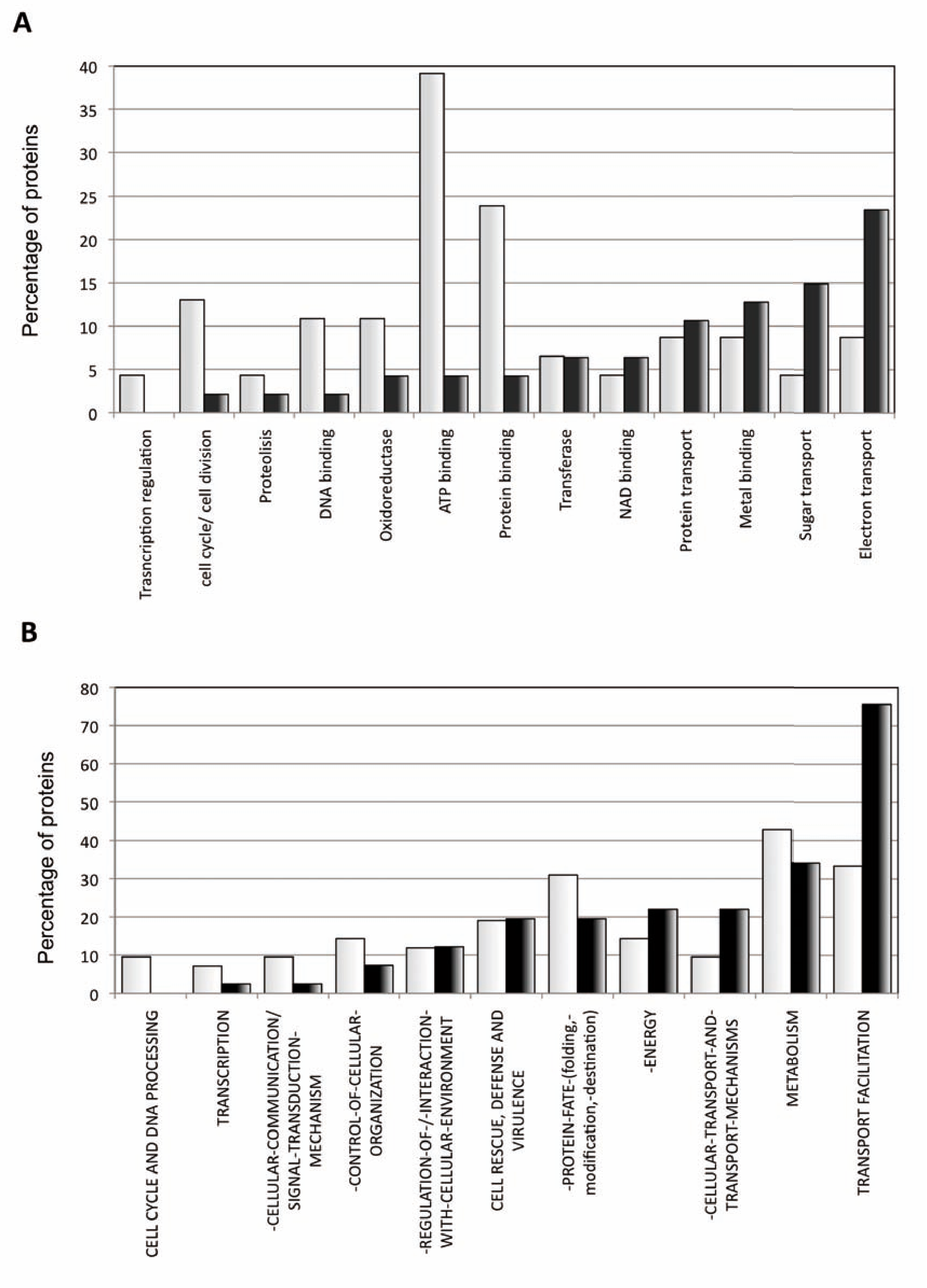

Number of predicted transmembrane segments
0
2
4
6
8
10
12
14
16
-41
-16
-12
-10
-6
-5
-3
-2
0
2
3
5
6
11
18
22
25
30
33
35
37
40
42
49
Na4vSS

Figure 11


A
Percentage of proteins
0
5
10
15
20
25
30
35
40
Trasncription regulation
cell cycle/ cell division
Proteolisis
DNA binding
Oxidoreductase
ATP binding
Protein binding
Transferase
NAD binding
Protein transport
Metal binding
Sugar transport
Electron transport
B
Percentage of proteins
0
10
20
30
40
50
60
70
80
CELL CYCLE AND DNA PROCESSING
TRANSCRIPTION
-CELLULAR-COMMUNICATION/ SIGNAL-TRANSDUCTION-MECHANISM
-CONTROL-OF-CELLULAR-ORGANIZATION
-REGULATION-OF-/-INTERACTION-WITH-CELLULAR-ENVIRONMENT
CELL RESCUE, DEFENSE AND VIRULENCE
-PROTEIN-FATE-(folding,-modification,-destination)
-ENERGY
-CELLULAR-TRANSPORT-AND-TRANSPORT-MECHANISMS
METABOLISM
TRANSPORT FACILITATION