Universitat Autònoma de Barcelona

Using Small Globular Proteins to Study Folding Stability and Aggregation

Virginia Castillo Cano

September 2012

Universitat Autònoma de Barcelona

Departament de Bioquímica i Biologia Molecular

Institut de Biotecnologia i de Biomedicina

Using Small Globular Proteins to Study Folding Stability and Aggregation

Doctoral thesis presented by Virginia Castillo Cano for the degree of PhD in Biochemistry, Molecular Biology and Biomedicine from the University Autonomous of Barcelona.

Thesis performed in the Department of Biochemistry and Molecular Biology, and Institute of Biotechnology and Biomedicine, supervised by Dr. Salvador Ventura Zamora.

Virginia Castillo Cano

Salvador Ventura Zamora

Cerdanyola del Vallès, September 2012

SUMMARY

The purpose of the thesis entitled “Using small globular proteins to study folding stability and aggregation” is to contribute to understand how globular proteins fold into their native, functional structures or, alternatively, misfold and aggregate into toxic assemblies.

Protein misfolding diseases include an important number of human disorders such as Parkinson’s and Alzheimer’s disease, which are related to conformational changes from soluble non-toxic to aggregated toxic species. Moreover, the over-expression of recombinant proteins usually leads to the accumulation of protein aggregates, being a major bottleneck in several biotechnological processes. Hence, the development of strategies to diminish or avoid these aberrant reactions has become an important issue in both biomedical and biotechnological industries.

In the present thesis we have used a battery of biophysical and computational techniques to analyze the folding, conformational stability and aggregation propensity of several globular proteins. The combination of experimental (*in vivo* and *in vitro*) and bioinformatic approaches has provided insights into the intrinsic and structural properties, including the presence of disulfide bonds and the quaternary structure, that modulate these processes under physiological conditions. Overall, the data illustrates how the establishment of native-like contacts providing folding intermediates, native structures or protein interfaces with significant thermodynamic stability is a crucial process both to drive protein folding and to compete toxic aggregation.

This thesis is based on the following scientific papers:

I. **Designing out disulfide bonds of leech carboxypeptidase inhibitor: implications for its folding, stability and function**. Arolas JL, Castillo V, Bronsoms S, Aviles FX, Ventura S. *J Mol Biol*. 2009 Sep 18; 392(2): 529-46.

II. **Deciphering the role of the thermodynamic and kinetic stabilities of SH3 domains on their aggregation inside bacteria**. Castillo V, Espargaró A, Gordo V, Vendrell J, Ventura S. *Proteomics*. 2010 Dec; 10(23): 4172-85.

III. **Amyloidogenic regions and interaction surfaces overlap in globular proteins related to conformational diseases**. Castillo V, Ventura S. *PLoS Comput Biol*. 2009 Aug; 5(8): e1000476.

Subsequent research works are included in the annex section:

IV. **The *in vivo* and *in vitro* aggregation properties of globular proteins correlate with their conformational stability: the SH3 case**. Espargaró A, Castillo V, de Groot NS, Ventura S. *J Mol Biol*. 2008 May 16; 378(5): 1116-31.

V. **The N-terminal helix controls the transition between the soluble and amyloid states of an FF domain**. Castillo V, Chiti F and Ventura S. *Submitted for publication*.

ABBREVIATIONS

1S	One-difulfide intermediates
2S	Two-difulfide intermediates
3S	Three-difulfide intermediates
4S	Four-difulfide intermediates
Å	Amstrong
Aß	Amyloid-ß peptide
ASA	Accessible surface area
ß2m	ß2 microglobulin
BPTI	Bovine pancreatic trypsin inhibitor
CD	Circular dichroism
CTD	Carboxyl-terminal domain
EGF	Epidermal growth factor
FALS	Familial amyotrophic lateral sclerosis
FTIR	Fourier transform infrared spectroscopy
GFP	Green fluorescent protein
IB	Inclusion body
LCI	Leech carboxypeptidase inhibitor
LDTI	Leech-derived trypsin inhibitor
mRNA	Messenger ribonucleic acid
N	Native state
NMR	Nuclear magnetic resonance
ODA	Optimal Docking Area
PDB	Protein Data Bank
RNase A	Ribonuclease A
SOD1	Superoxid dismutase 1
SPC-SH3	Spectrin-SH3
Sso AcP	*Sulfolobus solfataricus* acylphosphatase
TEM	Transmission electron microscopy
ThT	Thioflavin-T
TS	Transition state
TTR	Transthyretin
U	Unfolded state

INDEX

I - INTRODUCTION

1. - PROTEIN FOLDING

The process by which a polypeptide chain acquires its three dimensional native structure is denominated protein folding. After the amino acidic chain synthesis in the ribosome, local and long-range intramolecular interactions drive the folding process. The main forces leading to the attainment of the secondary and tertiary structure are hydrogen bonds, electrostatic interactions and hydrophobic forces. The native conformation is constituted by a large number of dynamic states, which display different but energetically proximal minimums. The fluctuations between these species provide flexibility, indispensable for protein functional activity under physiological conditions. Understanding the molecular basis of protein folding and unfolding has become a key topic in biology because mistakes in the acquisition or maintenance of the native state are being found linked to an increasing number of fatal diseases (Chiti and Dobson, 2006).

1.1 - First insights

In the early 1960s, Christian B. Anfinsen and co-workers studied the *in vitro* refolding of Ribonuclease A (RNase A), demonstrating that the amino acid sequence suffices to encode its three-dimensional structure (Anfinsen et al., 1961). The native state was proposed to be the most stable structure under physiological conditions and thermodynamics the main determinant of the folding process. Later on, Cyrus Levinthal illustrated how the acquisition of a particular structure by a random search would need astronomic time periods, considering the large number of possible conformations that the polypeptide should sample (Levinthal, 1968, Zwanzig et al., 1992). This argument leaded the search of folding pathways in the forthcoming years.

1.2 - Folding intermediates and folding pathways

Many efforts have been focused on the characterization of folding intermediates monitoring protein structural changes. Promoting jumps in the folding and unfolding reactions makes it possible to identify the folding model for a given protein. For instance, a protein follows a “two-state” model when only the unfolded or the native state are present at any point of the folding reaction. When partially folded

conformers are populated along the folding reaction, they can correspond both to on-pathway or off-pathway intermediates, depending if they can achieve the native structure or remain trapped in an energy minimum.

Several studies have revealed that folding pathways could be guided by specific native-like interactions directing polypeptide chains towards the acquisition of the lowest energetic structure (Daggett and Fersht, 2003). The “nucleation-growth” model proposed that secondary structures are propagated rapidly towards tertiary contacts leading to the acquisition of the native structure, but this mechanism didn’t take into account the presence of folding intermediates (Wetlaufer, 1973). The alternative “framework-model” postulated secondary structures as the first formed elements. The native form may be achieved then by docking the secondary structures (Kim and Baldwin, 1982, Kim and Baldwin, 1990, Karplus and Weaver, 1976). The last proposed model was that based on the “hydrophobic collapse”, in which hydrophobic forces promote a rapid compaction of the structure, reducing the conformational search towards the native structure (Baldwin, 1989). Finally, the “nucleation-condensation” model that combines the framework and the hydrophobic collapse mechanisms was proposed. A general hydrophobic collapse might promote the formation and stabilization of partially formed secondary and tertiary structures, leading to the formation of a collection of conformers named as the “molten globule” states (Fersht, 1995, Itzhaki et al., 1995, Fersht, 1997). This partly organized globular intermediate is observable at least in some proteins. Its acquisition is usually extremely fast (few milliseconds), which results in a structure less compact than of the native state. Therefore, the molten globule has to be considered as a collection of divergent structures which can rapidly interconvert. The unique functional form is reached slowly, involving the formation of native and proper packing interactions.

1.3 - Energy and kinetic diagrams

Overall, protein folding is a process controlled by thermodynamic and kinetic parameters. Interactions controlling the formation of secondary and tertiary structure, such as hydrogen bonds, electrostatic and hydrophobic forces turn folding into a thermodynamically favorable process. The free energy profiles show the acquisition of the native structure from its unfolded ensemble via a transition state ensemble,

resulting in a global decrease on the free energy of the system (figure 1A). A two-state system can be described by the following equation:

$$U \underset{k_U}{\overset{k_F}{\leftrightarrow}} N$$

where the equilibrium constant and the free energy associated are defined as;

$$k_{eq} = \frac{[N]}{[U]} = \frac{k_F}{k_U}; \quad \Delta G_{N-U} = -RT \ln k_{eq}$$

Entropic and enthalpic contributions determine the free energy of the system according to the following equation:

$$\Delta G_{N-U} = \Delta H - T\Delta S$$

The entropic contribution to folding is generally unfavorable because the attainment of a single ordered structure implies the loss of many initially unstructured conformations. However, the hydrophobic effect reduces the entropic penalty for folding, since water molecules, ordered around hydrophobic residues exposed to solvent in the unfolded ensemble, gain entropy upon folding.

Favorable enthalpic terms include the formation of stabilizing native interactions, such as electrostatic effects, hydrophobic forces, hydrogen bonds and van der Waals interactions.

The thermodynamic stability of a polypeptide is defined by the total energy difference between the native and the denatured state (ΔG_{N-U}) (figure 1A). Energetic barriers for folding (ΔG_{TS-U}) and unfolding (ΔG_{N-TS}) determine the free energy value, which usually ranges between 5 and 15 kcal/mol. This marginal stability is biologically very important, because living systems must regulate protein concentrations according to cellular requirements and several protein functions require structural flexibility. Proteins, under natural selection, are expected to have evolved an optimal stability to perform their function in cellular environments (Doig and Williams, 1992).

The time scale of protein folding may range from milliseconds to hours. In the framework of the Levinthal's paradox this means that the folding process must be directed in some way through a kinetic pathway to avoid a large number of irrelevant conformations. The folded state is kinetically trapped in a local free energy minimum, and separated from the unfolded ensemble through the unfolding activation barrier. Kinetic parameters associated to the folding process may be studied using the

stopped-flow technique, which allows monitoring unfolding and folding reactions and calculating their rates.

Thus, for two-state folding proteins the observed rate constant can be calculated as:

$$k_{obs} = k_F e^{(-m_F[D])} + k_U e^{(-m_U[D])}$$

where k_F and k_U are the rate constants at different denaturant concentrations ([D]) for folding and unfolding reactions, respectively. The logarithm of k_{obs} is assumed to depend linearly on the denaturant concentration, resulting in the slopes m_F and m_U, known as the kinetic m-values (figure 1B).

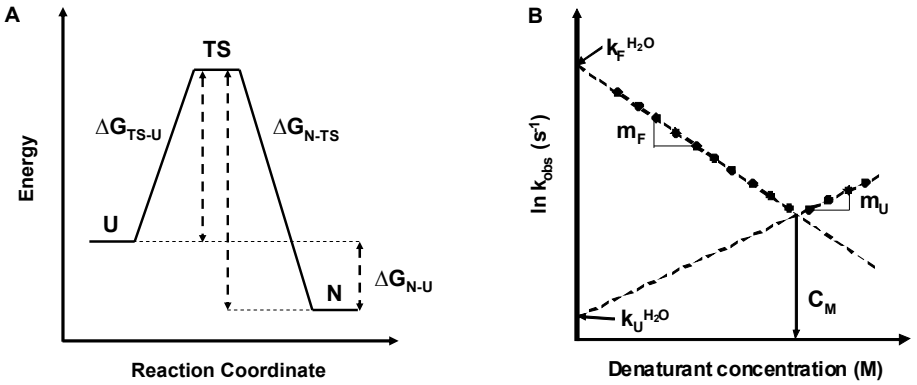


Figure 1. **A** Simplified energy diagram. The unfolded species (U) attain the native state (N) after reaching the transition state (TS), a maximum free energy state. The Gibbs energy (ΔG_{N-U}) is the difference between the energetic barrier for unfolding (ΔG_{N-TS}) and the energetic barrier for folding (ΔG_{TS-U}). **B** Schematic Chevron plot diagram. The circles represent the experimental k_{obs} at different concentrations of denaturant. The C_M value is the denaturation midpoint. k_F^{H2O} and k_U^{H2O} are the folding and unfolding rates in the absence of denaturant, respectively. The slopes of the respective folding and unfolding regions are the known kinetic m-values: m_F and m_U.

1.4 - Energy landscapes

During protein folding, polypeptide chains adopt a large number of different conformations since, apart from backbone flexibility, each residue is able to adopt different dihedral angles in unfolded or partially folded conformations. The ensemble of structures populating the folding reaction can be represented using energy landscapes (figure 2). These three dimensional diagrams consider the energies and entropies of multiple states, showing the most thermodynamically stable or metastable conformations as energetic minimums in the lanscape.

The establishment of specific native-like intramolecular contacts canalizes the formation of on-pathway partially folded intermediates and thus the attainment of the native structure. By contrast, the formation of non-native intermolecular interactions can promote the appearance of amorphous and fibrillar aggregates, corresponding to competing strong energetic minima.

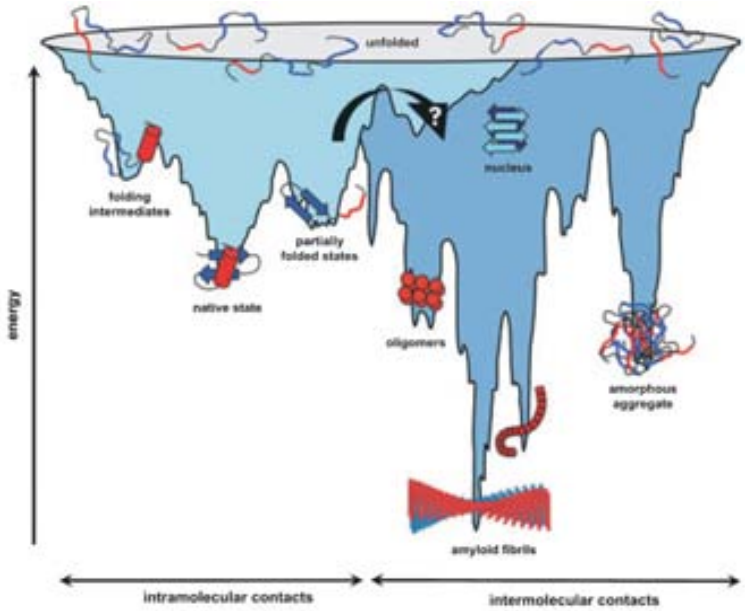


Figure 2. Energy landscape for protein folding and aggregation. Driving forces towards the acquisition of the native state are intramolecular interactions, whereas intermolecular contacts allow the formation of toxic species, achieving thermodynamic minimums that are more stable than the native structure (Jahn and Radford, 2005).

The shape of energy landscapes depend on the polypeptide sequence and environmental factors. For small proteins, these landscapes are funnel-like since they have a reduced number of denatured conformations promoting an efficient and rapid acquisition of the native structure through unstable transitions states of higher energy. By contrast, larger proteins acquire a wide range of non-native populating species, which can achieve the folded state through metastable intermediate states with local low energy minima.

Stopped-flow approaches combined with computer simulations have allowed studying the transition states of folding of different model proteins (figure 3). Usually, transition states show a native-like fold where most of the hydrophobic residues are buried in the core of the structure. The acquisition of the transition state represents

the rate-limiting step, since a relative large energetic barrier must be overcome. Then, the folded state should be easily reached (Dinner and Karplus, 1998).

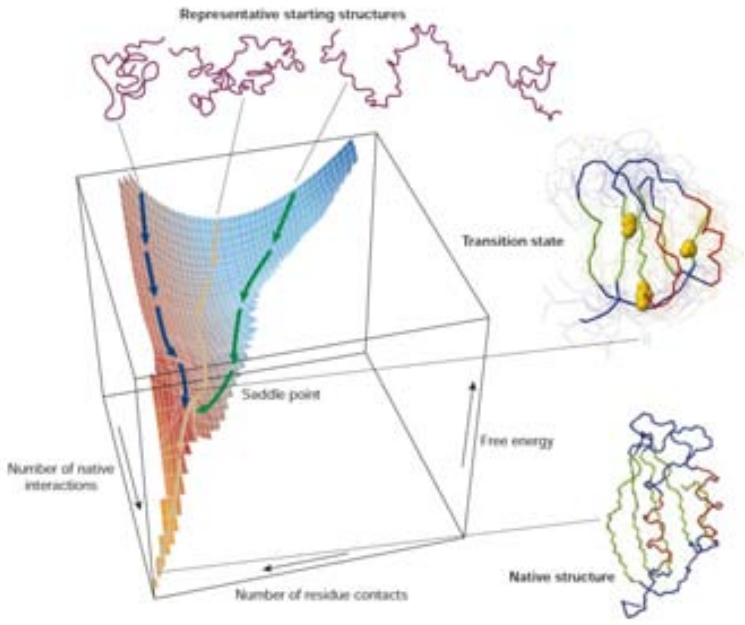


Figure 3. Energy landscape derived form computer simulation of a simplified model of a small protein. The multiple denatured conformations are funneled towards the acquisition of the native structure. In this case, the transition state is the barrier that almost all the proteins must overcome to achieve the native fold (Dobson, 2003).

1.5 - Disulfide bonds and protein folding

The role of disulfide bridges in protein folding has been largely studied using the RNase A as a disulfide protein model. The process by which a protein attains its native disulfide bonds (*disulfide-bond regeneration*) and its native structure (*conformational folding*) is named oxidative folding. However, pairing between cysteines is a complex process where oxidation, reduction and *disulfide reshuffling* (when a disulfide bond is attacked by a thiolate group of the same protein) might compete inside the cells (Wedemeyer et al., 2000). In eukaryotic cells, disulfide bond formation occurs in the endoplasmatic recticulum before the transport of most extracellular and membrane-bond proteins. Intracellular disulfide-containing proteins are rarer because the cytosol tends to be a reducing environment. Disulfide formation *in vivo* is catalyzed by specialized enzymes, such as the protein disulfide isomerase (PDI), which catalyzes internal disulfide exchange to remove scrambled intermediates with incorrectly bonded disulfide bridges.

This covalent interaction offers several important advantages for protein folding reactions (Wedemeyer et al., 2000). First, the oxidation and reduction of a disulfide bridge follows a two-state mechanism and is a localized and structural change in the polypeptide chain. Second, disulfide bonds stabilize significantly proteins, because the conformational fluctuations of the unfolded ensemble and the associate entropy are strongly reduced (Poland and Scheraga, 1965, Pace et al., 1988, Arolas et al., 2005b, Lin et al., 1984). Actually, a good strategy to increase protein stability is crosslinking sequentially distant regions of a polypeptide chain (Grana-Montes et al.). The stability provided by disulfides is especially important for protecting secreted proteins from the oxidant and proteolytic external environment, thus preventing their inactivation and increasing their half-life (Zavodszky et al., 2001). And third, the folded state is in many cases enthalpically stabilized through favorable local interactions, for instance by stabilizing the packing of a local cluster of hydrophobic residues. Disulfide bonds are commonly buried in local hydrophobic regions protecting them from redox reagents and thiolate groups of the protein itself, which might suggest a strong cooperativity between local and global folding.

Overall, disulfide bridges provoke well-known chemical and structural changes, allowing us to use them as good reporters of folding pathways. Oxidation and reduction reactions are useful strategies to study disulfide folding mechanisms, since folding intermediates can be chemically trapped by alkylation or acidification, and structurally characterized (Creighton, 1990, Narayan et al., 2000). The irreversibly alkylation approach usually promotes harmful steric effects on the conformation of folding intermediates. On the contrary, quenching by acidification has become the standard trapping method, because of the rapidly isolation of unmodified intermediates. The subsequent separation procedure includes many chromatographic and electrophoretic methods. The most useful technique for separating acid-trapped intermediates is the reversed-phase high-performance liquid chromatography (RP-HPLC). When chemical quenchers are used, the increase in molecular weight can be monitored by mass spectrometry (MS). Finally, disulfide pairing of the intermediates is analyzed by digestion with endopeptidases that allows obtaining the peptide mass fingerprints.

In oxidative folding experiments *in vitro*, the protein is placed under denaturing and reducing conditions. Then, the sequential disulfide-bonds formation is promoted in removing both chaotropic and reducing agents. Monitoring disulfides regeneration can be performed in the absence or presence of redox agents (Chatrenet and Chang, 1993). Oxidative and reductive reagents increase the efficiency of the oxidative folding process, and promote disulfide reshuffling allowing intermediates to attain the native disulfide connectivity. Different redox conditions can be used to modify refolding rates and folding efficiencies, remaining the folding pathways unaffected, as demonstrated in the case of hirudin protein (Chang, 1994).

Reductive unfolding assays are used to provide information about the stability of disulfide bonds and their degree of exposition to the reductive environment. Native-disulfide proteins are treated with different concentrations of reducing agent, allowing us to calculate thermodynamic values from their reductive rates. *Disulfide scrambling* is a recent methodology allowing determination of stability toward chemical denaturants that, in the presence of a thiol reagent, promotes the appearance of a variety of scrambled species which contain the same number of native disulfide bonds but mainly non-natively connected. The called “des species” are intermediates that lack one disulfide bond and comprise native disulfide-bond pairings and native-like structures.

The use of stop/go folding experiments offers the opportunity to examine and compare the properties of isolated intermediates (Chang, 1993). This approach also allows evaluation of their kinetic roles in folding pathway. Importantly, disulfide folding intermediates can be characterized structurally because in most cases, are stable and can be isolated.

1.6 - Folding pathways of small disulfide proteins

The folding and unfolding pathways of a large number of disulfide-rich proteins have been characterized (Arolas et al., 2006b, Mamathambika and Bardwell, 2008). Surprisingly, a high diversity of folding pathways exist, especially for small disulfide-containing proteins (Arolas et al., 2006a, Chang and Ventura, 2011) (figure 4).

These data reveals how difficult is to predict the folding pathways from the primary sequence, as in the cases of tick anticoagulant peptide (TAP) and bovine

pancreatic trypsin inhibitor (BPTI), two proteins displaying the same fold and with the same disulfide connectivity but exhibiting stricking different oxidative folding pathways (Weissman and Kim, 1991, Chang, 1996, Chang and Li, 2005).

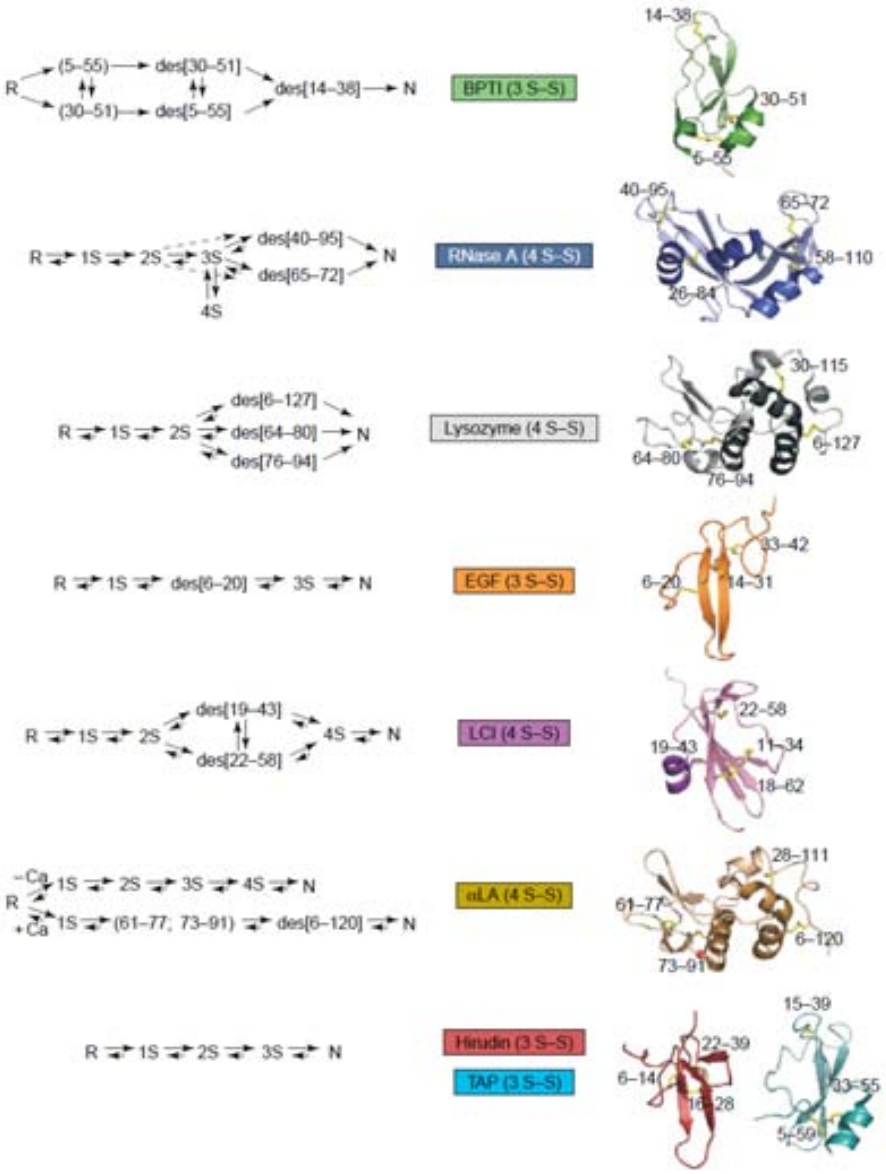


Figure 4. Comparative folding of small disulfide-rich proteins. R and N indicate the reduced and native forms, respectively. 1S, 2S, 3S and 4S are ensembles of molecules with the corresponding number of disulfide bonds. "Des species" are indicated with the prefix "des" and the missing disulfides are between parentheses (Arolas et al., 2006a).

Oxidative folding pathways are characterized by: (1) the heterogeneity of intermediate species populating the folding reaction; (2) the frequency of folding intermediates containing native disulfide bonds and native-like structure; and (3) the

presence of fully oxidized scrambled species displaying two non-native disulfide bonds. Two extreme models exist: proteins that fold through folding intermediates containing exclusively native disulfides; and proteins that fold through fully oxidized scrambled isomers or mostly non-native disulfide intermediates (Chang, 2004, Chang, 2008).

BPTI is the first and the best characterized model of small disulfide-rich protein. Its folding pathway contains a limited number of intermediates with mostly native-like structures and native disulfide connectivity. (Creighton, 1990, Creighton and Goldenberg, 1984, Weissman and Kim, 1991, Weissman and Kim, 1992, Dadlez, 1997). The folding behavior of proteins sharing this property is described as BPTI-like. This is the case of leech-derived trypsin inhibitor (LDTI), a protein that binds tightly to a human trypsin-like serine proteinase involved in allergic and inflammatory diseases. This 46-residue protein contains three disulfide bonds and folds through a sequential oxidation of its cysteine residues. Similar to the BPTI case, only five of the 74 different possible intermediates are accumulated in the folding pathway (Arolas et al., 2008). The initial oxidation of one of the LDTI disulfides provides an ensemble of one-difulfide (1S) intermediates that preferably contains a native disulfide bond linked to secondary structure elements, β1- and β3-strand (Pantoja-Uceda et al., 2009). The further oxidation (of 1S intermediates) results in only three 2S folding species, all of them presenting native disulfide bonds. Surprisingly, only one of the three 2S is able to reach the native structure. This data suggest a tight association between the formation of native disulfides and conformational folding in BPTI-like proteins. Moreover, the simplicity of the conformational landscape provokes an efficient and fast folding process.

In contrast to LDTI folding pathway, hirudin protein is characterized by folding through scrambled isomers, without any preferential accumulation of folding intermediates (Chatrenet and Chang, 1992, Chatrenet and Chang, 1993, Chang, 1994). This thrombin inhibitor 64-residue protein folds into two structurally different domains: the N-terminal globular domain, which contains three disulfide bonds; and the disordered C-terminal domain (Folkers et al., 1989, Rydel et al., 1990, Grutter et al., 1990). Hirudin-like proteins are characterized by an initial unspecific packing of the structure, and a final stage in which the collapsed structures acquire the native structure. The first step involves a sequential oxidation of hirudin-free cysteines and

multiple disulfide reshuffling reactions resulting in 1S, 2S and 3S-scrambled isomers. Mainly 3S-scrambled intermediates with at least two non-native disulfides are accumulated in the folding reaction. Differing from BPTI-like proteins, the oxidative folding of hirudin-like proteins follows apparently a “trial-and-error” process where scrambled isomers act as major kinetic traps, resulting in a slow and inefficient folding mechanism.

The epidermal growth factor (EGF) is one of several proteins displaying both BPTI-like and hirudin-like mechanisms. EGF is a 53-residue protein containing three disulfide bonds that stabilize three different loops (Montelione et al., 1992, Ogiso et al., 2002). The initial step of its folding pathway comprises the formation of a heterogeneous ensemble of 1S intermediates (Chang et al., 2001), which transforms rapidly into a single and predominant two native-disulfides intermediate. One should expect it to be easy the further oxidation of the third native disulfide, however, scrambled isomers are an obligatory step before reaching the native functional form. This last step constitutes the major rate-limiting step, decreasing the efficiency and velocity of the folding reaction. Similar to EGF, the oxidative folding pathway of leech carboxypeptidase inhibitor (LCI) undergoes a sequential flow through 1- and 2-disulfide intermediates that rapidly accumulate as two predominant 3-disulfide intermediates with native disulfide pairings. These two species act as the major kinetic traps, which need structural rearrangements through the formation of 4-disulfide scrambled isomers to attain the native structure (Salamanca et al., 2003, Arolas et al., 2004).

2. - PROTEIN MISFOLDING AND DISEASE

During the folding process, failures in the acquisition or maintenance of the native structure have important consequences, in many cases leading to cellular toxicity. These pathological conditions are generally named protein misfolding (or protein conformational) diseases (Chiti and Dobson, 2006).

Partially folded species expose to the solvent regions that are buried in their native states, which allow the establishment of non-native intermolecular interactions leading to protein deposition. Organisms have developed quality control mechanisms

to prevent misfolding, such as chaperone-assisted folding or degradation through ubiquitin-proteasome pathway. Generally, failures in these strategies lead to dramatic effects for living organisms. More than forty human conformational diseases are related with the formation of intracellular and/or extracellular amyloid-like aggregates. These disorders can be classified into three groups depending on the localization of the aggregates: neurodegenerative amyloidoses, in which aggregation occurs in the brain (Parkinson's and Alzheimer's disease); non-neuropathic localized amyloidoses, in which aggregation occurs in a single type of tissue other than the brain (tipus II diabetes); and non-neuropathic systemic amyloidoses, when aggregation occurs in multiple tissues (Lysozyme and Amiloyd Light-chain amyloidoses) (Chiti and Dobson, 2006). The origin of these illnesses is mainly sporadic (85%) and less frequently hereditary (10%). However, it is known that 5 % of these diseases can be transmissible between mammals, as spongiform encephalopathies.

3. - PROTEIN AGGREGATION AND AMYLOID FIBRILS FORMATION

3.1 - Amyloid fibrils

Aggregating proteins and peptides involved in conformational diseases display different structural states in solution, ranging from globular to totally unstructured conformations. However, the fibrillar structures characteristic of many of their aggregates share common morphological features. For instance, amyloid fibrils present a typical "cross-β" X-ray fiber diffraction pattern corresponding to highly ordered structures where intermolecular β-sheets are perpendicular to the fibril axis (figure 5). The molecular details of fibrillar structures can be accessed using High-Resolution Solid-State nuclear magnetic resonance (ssNMR) (Petkova et al., 2002, Ritter et al., 2005). Transmission electron microscopy (TEM) and atomic force microscopy (AFM) allow to visualize fibrillar aggregates *in vitro,* revealing that fibrils are usually composed by a group (2-6) of protofilaments each about 2-5 nm of diameter. Protofibrillar structures with spherical conformation have been also observed for several proteins such amylin (Kayed et al., 2004), β2-microglobulin (β2m) (Gosal et al., 2005) and acylphosphatase from *Sulfolobus solfataricus* (*Sso* AcP) (Plakoutsi et al.,

2004). Besides, amyloid-like aggregates have the capability to bind specific dyes such Thioflavin-T (ThT), increasing its intrinsic fluorescence, and Congo red (CR), promoting a red-shift in its absorption spectrum.

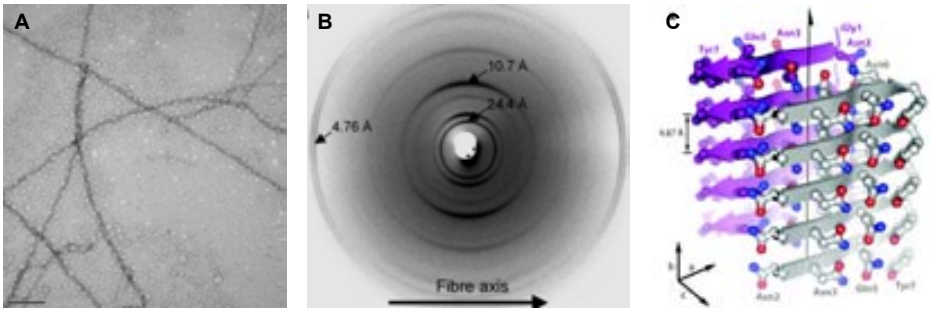


Figure 5. **A** Electron micrograph of URN1-FF amyloid fibrils. **B** X-ray diffraction pattern of SUP35 fibres, adapted from Makin 2005. **C** GNNQQNY amyloid structure obtained by X-ray diffraction, ribbon diagram showing the pair of cross-β sheet spine (backbones are showed in arrows and sidechains by ball-and-stick structures, PDB 1YJP) (Nelson et al., 2005).

3.2 - Functional fibrils

The conversion of some polypeptide chains into fibrillar species has functional purposes in living systems. Curlin protein of *Escherichia coli* is used to colonize inert surfaces and mediate binding to host proteins (Barnhart and Chapman, 2006, Chapman et al., 2002). Another example is the filamentous bacterium *Streptomyces coelicolor* which produces aerial hyphae leading to the efficiently dispersion of its spores. Chaplins proteins have been identified in these hyphae which have the ability to form amyloid fibrils acting cooperatively to allow the formation of aerial structures (Claessen et al., 2003).

Other proteins can shift their soluble state to fibrillar structures conferring specific functional roles inside the cells. Some prion proteins are composed by a globular domain and an unstructured fragment leading the conversion between both states. This transition can be associated with important phenotype changes as in the case of Ure2p (Chien et al., 2004, Nakayashiki et al., 2005) and Sup35 (Eaglestone et al., 1999, True and Lindquist, 2000), prion proteins from *Saccharomyces cerevisiae*. The polymerization of the HET-s prion protein from *Podospora anserine* is involved in a controlled programmed cell death phenomenon (Coustou et al., 1997, Saupe, 2000).

3.3 - Fibril formation mechanism

Polypeptide chains adopt a large number of conformations during multi-step amyloid self-assembling processes. The precursors must expose aggregation-prone regions to allow the establishment of intermolecular contacts. Thus, native globular proteins need to be destabilized by mutations or external factors, such as high temperature, changes on the pH or addition of denaturants (Chiti et al., 1999). Nevertheless, recent studies have revealed that native-like conformations as well as fragments of proteins generated by proteolysis unable to fold are capable to initiate the protein deposition mechanism (Sabate et al., 2010 Chiti and Dobson, 2009). Overall, the ensemble of precursors promoting protein deposition is really heterogeneous, such as degraded fragments, totally or partially unstructured species or native-like structures, which coexist in equilibrium under physiological conditions (figure 6).

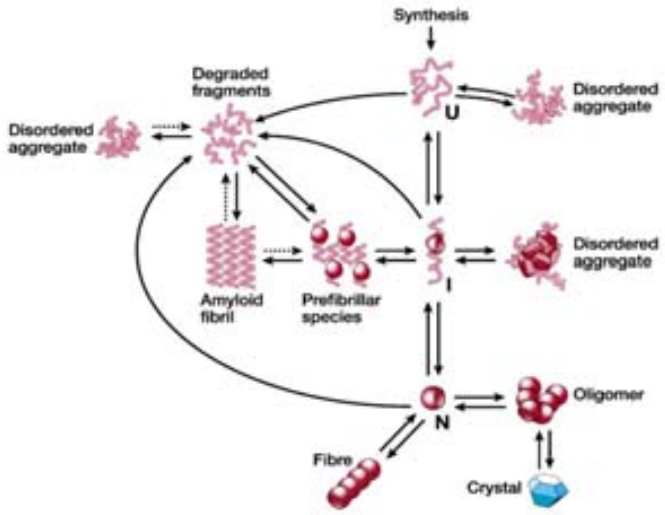


Figure 6. Schematic representation of the conformational diversity that polypeptide chains might adopt. New synthesized proteins can fold into the native structure. Nevertheless, degradation and aggregation (including amyloid fibrils and functional oligomers) can compete under physiological conditions. The interconversion between these populations is determined by their thermodynamic and kinetic stability (Dobson, 2004).

For a two-state folding protein, the aggregation mechanism is assumed to be an irreversible process occurring only from unfolded species. The following equation describes this behavior, known as the Lumry-Eyring model (Sanchez-Ruiz, 1992):

$$F \underset{k_F}{\overset{k_U}{\leftrightarrow}} U \overset{k_A}{\rightarrow} A$$

k_F, k_U and k_A are the folding, unfolding and aggregation rates, respectively.

The native ensemble is separated from non-native species by kinetic barriers, especially the unfolding activation barrier. Thus, unfolding process is considered the rate-limiting step because the deposition mechanism can not be faster than the unfolding rates. According to this hypothesis, thermodynamic stability does not guarantee that a protein will remain in the native state during a given timescale. Indeed, many proteins have been designed by evolution to have high activation barriers for unfolding (kinetic stability), thus balancing the irreversibly effect of aggregation (Plaza del Pino et al., 2000).

The conversion of soluble monomers into amyloid fibrils may be followed by ThT fluorescence, light scattering and circular dichroism (CD) spectroscopy (Naiki et al., 1997, Serio et al., 2000, Uversky et al., 2002, Pedersen et al., 2004). Initially, structurally diverse precursors promote the nucleation process, the slowest and protein-concentration dependent step, as indeed for crystallization. Protein monomers need to transform their conformation to β-sheet structure acquiring a specific critical size, and resulting in a thermodynamically unfavorable process. The nucleation stage is visualized as a lag phase (figure 7), which can be temporarily reduced by the addition of preformed prefibrillar aggregates to the aggregation assay, known as seeds (Chiti and Dobson, 2006).

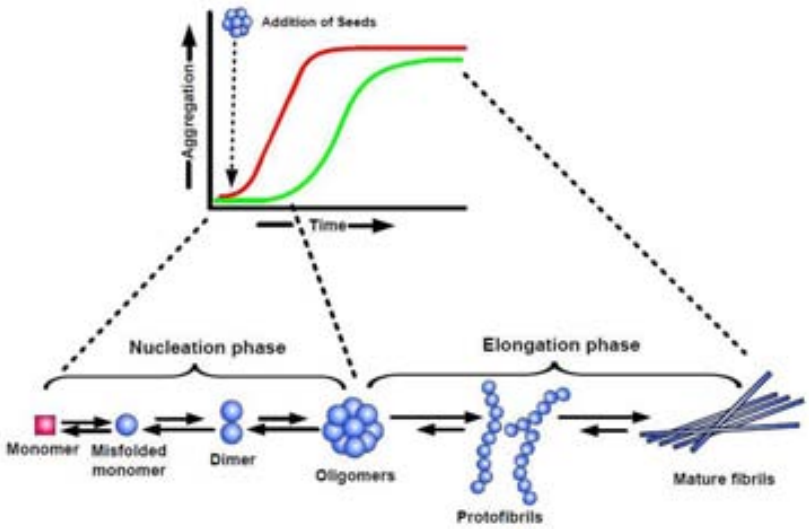


Figure 7. Nucleation-polymerization model of amyloid aggregation. The first phase is the nucleation lag phase, in which monomers undergo conformational change and associate to form oligomeric nuclei, a thermodynamically unfavourable process. The second phase is the elongation/growth phase, in which the nuclei rapidly grow by further addition of monomers and form larger fibrils, which is a fastly and favourable process. The sigmoidal curve shows the lag phase (rate limiting step) followed by a rapid growth phase. Thus, the addition of preformed seeds reduces the lag time (Kumar and Walter, 2011).

After the achievement of the oligomeric species that form the aggregation nucleus, monomers can be incorporated rapidly and efficiently during the fibril elongation process. This growth phase is thermodynamically favorable and usually corresponds to a single-exponential curve.

3.4 - Native-like aggregation

Several globular proteins have been shown *in vitro* to undergo amyloid fibril formation under solution conditions that promote their partial unfolding. Nevertheless, native-like conformations without the requirement to cross the major energy barrier for unfolding are also in certain cases able to initiate the self-assemble reaction (Bemporad and Chiti, 2009) (figure 8). Essentially, these conformational states are thermodynamically different from the native state, and can be accessed directly from the native state through thermal fluctuations or changes on the physiological pH (Bemporad and Chiti, 2009). Structural fluctuations can be also facilitated by mutations, which cause the unfolding of structured regions or by dissociation of the quaternary structure (Chiti and Dobson, 2009).

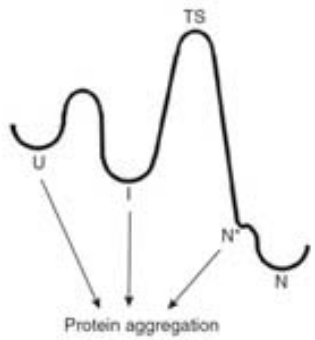


Figure 8. Energy diagram for protein folding. According to the classical thermodynamic and kinetic principles, unfolded species (U) can attain the native form (N) crossing the major free energy barrier of their transition state (TS). Other locally unfolded states (N*) with higher energy states than N might become accessible from N via thermal fluctuations. Polypeptide chains adopting U, I and N* states are able to self-assemble and trigger amyloid formation, being N* easier to access from N than are I and U (Chiti and Dobson, 2009).

A well studied protein model for native state amyloid aggregation is the *Sso* AcP (Bemporad and Chiti, 2009). This α/β enzyme contains an unstructured N-terminal segment (Corazza et al., 2006), which plays a crucial role during the aggregation process. Many evidences have demonstrated that *Sso* AcP adopts native-like conformations under specific aggregating conditions, as indicated by far- and near-UV CD, enzymatic activity assays and stopped-flow measurements of folding and unfolding rates (Soldi et al., 2005, Plakoutsi et al., 2005, Plakoutsi et al., 2006, Bemporad et al.,

2008) (figure 9). In the first step of aggregation, initial native conformations convert into small oligomeric species, promoted by interactions between the unstructured N-terminal segment and a peripheral β-strand of the globular domain. These species present native-like conformation, but do not yet bind to specific amyloid dyes (Plakoutsi et al., 2005). In the second step, oligomeric species transform into spherical structures and chain-like protofibrils, showing a significant β-sheet structure and positive binding to amyloid dyes. Interestingly, aggregation rates for this protein are really faster than the unfolding ones, suggesting that global unfolding is not a necessary condition to initiate the aggregation process (Plakoutsi et al., 2004).

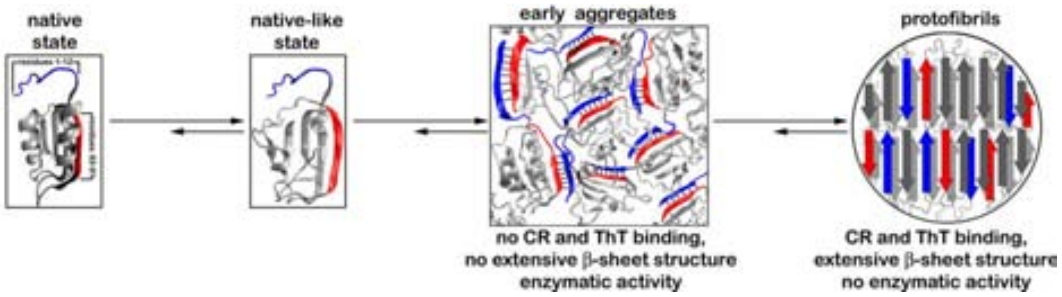


Figure 9. Proposed process of aggregation for *Sso* AcP. Initially, the polypeptide chain folds in its native monomeric form (N). Native-like conformations (N*) can be accessible through thermal fluctuations or destabilizing mutations. This transition is not a global unfolding process across the major free energy barrier. The formation of native-like aggregates is established trough the interaction between the unstructured N-terminal segment and a portion of the globular structure (the peripheral β-strand 4). Finally, these aggregates allow the convertion into amyloid-like protofibrils (Bemporad and Chiti, 2009).

In vitro studies of disease-related single mutants of human lysozyme protein exhibit local cooperative unfolding under physiological conditions. This transition does not involve crossing the energy barrier for unfolding, since partially unfolded species are accessed at least five orders of magnitude faster than overall unfolding (Canet et al., 2002, Dumoulin et al., 2005).

In the case of a superoxid dismutase 1 (SOD1) variant associated with familial amyotrophic lateral sclerosis (FALS), a highly flexible loop plays a key role in aggregation. This SOD1 mutant displays native-like dimeric structure, but temporary soluble oligomers can be formed through interactions involving the unstructured loop (Elam et al., 2003, Banci et al., 2005).

This behavior has been also found in other proteins such as transthyretin (TTR). Native TTR homotetramer needs to dissociate into a partially unfolded monomeric states, which are able to initiate the amyloid fibrillar process (Colon and Kelly, 1992, Lai et al., 1996, Quintas et al., 2001). TTR-associated familial diseases are linked to mutations that favor the aggregation mechanism (Sekijima et al., 2005). In contrast to SOD1 mutants, the unfolded regions of TTR monomeric species do not mediate the interactions between the monomers in the fibrils, but rather these unstructured regions remain exposed to the solvent in the fibrils (Serag et al., 2002).

Amyloid fibril formation of β2m protein has been found in dialysis-related amyloidosis. During the folding reaction, it has been identified at least one intermediate that accumulates after the major energy barrier for folding (Chiti et al., 2001, Jahn et al., 2006). Structural studies have shown native-like compactness and secondary structure for this state (Kameda et al., 2005).

4. - *IN VIVO* PROTEIN AGGREGATION

The formation of insoluble protein deposits causes a high number of devastating human diseases of growing incidence such as Alzheimer's disease, Parkinson's disease and type II diabetes among others. Amyloid deposits can be formed inside and outside the cell, involving predominantly the aggregation of a specific protein although other molecules can be incorporated, such as glycosaminoglycans and apolipoproteins (Westermark et al., 2007).

During protein translation, emerging polypeptide molecules are extremely susceptible to aggregation in the crowded cell environment (figure 10). Usually, the elongation reaction is faster than their folding, increasing the presence of totally or partially unfolded species inside the cells. The maintenance of the native structure depends on the balance between the protein folding rate and conformational stability. To guarantee the correct folding of polypeptide chains, cells have developed quality control systems constituted by molecular chaperones and proteases. These machineries recognize and interact with soluble, unfolded and misfolded polypeptide chains performing the refolding or degradation of these species (Hartl and Hayer-Hartl,

2002, Wickner et al., 1999). Interestingly, recent data shows that some chaperones are able to re-dissolve protein aggregates (Lee et al., 2004).

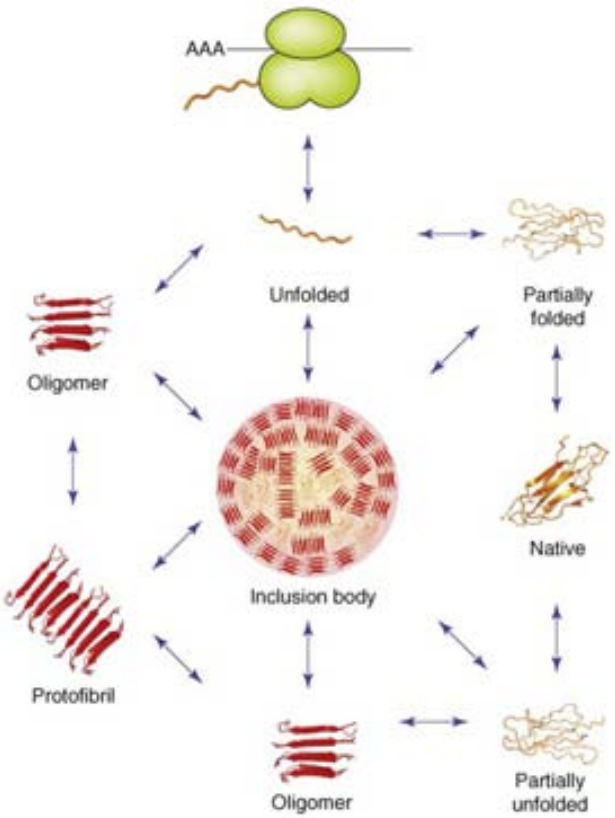


Figure 10. Promoting protein conformations of IB formation. New synthetized polypeptide chains in the ribosome are supposed to fold into soluble native conformations. However, during protein over-expression totally and partially unfolded species often establish anomalous but specific interactions, forming the aggregating nuclei. The self-assemble process of both oligomeric and native-like species promotes β-sheet enrichment in the IB structure. Conformational fluctuations and local unfolding allow the interconversion of the species between the soluble and insoluble fraction inside the cells (de Groot et al., 2009).

Prokaryotic organisms have been extensively used as simple cellular models, providing significant features about several biological processes such as aging or *in vivo* protein deposition (Garcia-Fruitos et al., 2005a, Gonzalez-Montalban et al., 2007, Sabate et al., 2010, Lindner et al., 2008). Many biotechnological processes employ *Escherichia coli* as a rapid and economical system in the over-expression of recombinant proteins. Generally, over-expression of recombinant polypeptides promotes the formation of protein aggregates, known as inclusion bodies (IBs).

Aggregation inside bacteria has been regarded as an unspecific process of non-structured and inactive protein deposits. However, high-resolution techniques have allowed studying in detail the inner structure of IBs, demonstrating that this *in vivo*

deposits share common structural properties with pathogenic amyloid aggregates in higher organisms (Morell et al., 2008, Wang et al., 2008). Their amyloid nature can be studied by electron microscopy and binding to amyloid specific dyes (Morell et al., 2008, Carrio and Villaverde, 2005). Additionally, recent studies have demonstrated the ability of amyloid-β peptide (Aβ) IBs to seed the fibrillar process (Morell et al., 2008).

These intracellular aggregates are dense, porous, hydrated and apparently amorphous structures of about 1 μm of diameter (Arie et al., 2006, Carrio et al., 2000). Mature IBs are usually composed by more than 90 % of recombinant protein but also may contain traces of molecular chaperones (Carrio et al., 2000, Carrio et al., 1998, Carrio and Villaverde, 2002, Schrodel and de Marco, 2005). Partially active conformations and native-like species can be found within bacterial aggregates, which are commonly coexisting with an extent of β-sheet structure (de Groot and Ventura, 2006, Garcia-Fruitos et al., 2005b, Oberg et al., 1994). Thus, specific interactions between solvent-exposed hydrophobic regions do not necessarily disturb the conformation of all protein domains, which allows maintaining in some cases the enzymatic activity. The green fluorescence protein (GFP) fused to a set of Aβ variants has been extensively used to characterize *in vivo* folding kinetics. Fast aggregating variants did not produce fluorescent IBs, since GFP protein is unable to fold correctly. In contrast, highly fluorescent aggregates were obtained for slow aggregating sequences (de Groot et al., 2006), illustrating a kinetic competition between folding and aggregation inside the cell.

5. - MOLECULAR DETERMINANTS AND PREDICTION OF AMYLOID AGGREGATION

Despite it was thought that the ability to form amyloid-like aggregates was restricted to proteins involved in amyloidoses, an increasing number of globular proteins not related to disease have been shown to form amyloid fibrils in the last years. These proteins do not share any sequential or structural similarity. In this way, aggregation might be proposed as a generic and intrinsic property of polypeptides in both eukaryotic and prokaryotic organisms (de Groot et al., 2009, Chiti and Dobson, 2006, Chiti et al., 2003).

5.1 - Intrinsic physico-chemical properties and external conditions

Protein aggregation processes depend on both intrinsic polypeptide properties and environmental conditions (DuBay et al., 2004). Intrinsic factors comprise several characteristics of the amino acid residues, such as hydrophobicity, charge, polarity and propensity to adopt secondary structure motifs. For many years, hydrophobic interactions have been considered as the main force in protein folding and aggregation (Fink, 1998). Actually, higher aggregation propensities have been associated with an increase in the hydrophobicity of polypeptides (Chiti et al., 2003). On the contrary, inverse correlation has been observed between net polypeptide charge and their aggregation propensity (Chiti et al., 2003, DuBay et al., 2004), because of the electrostatic repulsion promoted by the net charge during the establishment of the first intermolecular interactions. In the case of globular proteins, thermodynamic stability of the native state is often inversely correlated with the aggregation propensity (Hurle et al., 1994, Quintas et al., 1999, Chiti et al., 2000).

Environmental factors are also important in the tendency of polypeptides to form amyloid structures. Extrinsic properties that modulate protein aggregation include interactions with cellular components such as chaperones, proteases and the ubiquitin-proteasome system. Besides, physico-chemical parameters as pH, temperature, ionic strength and protein concentration modulate protein deposition. The morphology of amyloid fibrils can be remarkably diverse and influenced by the factors related above. For instance, salt concentration can strongly modulate the kinetics (mainly the length of the lag phase) and structure of amyloid fibrils of a SH3 mutant domain of α-spectrin (Morel et al., 2010). Low salt concentration promotes the slow formation of well-ordered and twisted fibrillar structures. In contrast, thinner and shorter fibrils are rapidly formed under high salt concentrations and high temperatures. Changes on the pH also have consequences in the deposition process, increasing the aggregation rates at low pH.

5.2 - Hot spots and gatekeepers

The primary sequence strongly influences the aggregation propensity, indeed point mutations may have a huge impact on protein solubility (Chiti et al., 2003). When aggregation is initiated from totally unfolded species, a high correlation is observed

between the aggregation rates resulting from single amino acid mutations and the effect of the substitutions on the intrinsic properties described above (Chiti et al., 2003). Not all the residues of polypeptide chains have the same contribution to the aggregation propensity. The regions that direct the deposition process have been called “hot spots” (Ventura et al., 2004, Ivanova et al., 2004), and they correspond to specific regions enriched in aliphatic and aromatic residues (Val, Leu, Ile, Phe, Tyr, Trp) (Monsellier et al., 2008, Rousseau et al., 2006). The presence of these specific regions in globular proteins is expected since the biophysical rules that promote the formation of native contacts and β-sheet intermolecular interactions are similar (Routledge et al., 2009). Evolutionary pressure against aggregation has employed amino acid substitutions in the flanks of hot spot regions (Monsellier and Chiti, 2007, Monsellier et al., 2007). Essentially, these so-called “gatekeepers” use the repulsive effect of their charge such as Arg and Lys residues, and amino acids incompatible with a β-structure (Pro), which are able to modulate the aggregation propensity of the sequences they flank (Rousseau et al., 2006, Reumers et al., 2009, Otzen et al., 2000, Richardson and Richardson, 2002, Monsellier et al., 2008). Small sequential changes either inside or close to these specific regions strongly affect protein solubility.

5.3 - Aggregation prediction algorithms

The ability to predict amyloid fibril formation is important for several reasons. The increasing knowledge about the structural features of amyloid fribrils and the forces that promote and stabilize their formation has allowed developing several mathematical tools. These prediction algorithms are able to detect aggregation-promoting regions and forecast protein aggregation propensities. More than ten different predictive tools have been developed, based on different empirical or structural parameters (Belli et al., 2011, Castillo et al., 2011). Empirical tools try to relate *in vivo* and *in vitro* experimental results with amino acid intrinsic properties (Chiti et al., 2003, DuBay et al., 2004, Tartaglia et al., 2004, Conchillo-Sole et al., 2007, Zibaee et al., 2007, Tartaglia and Vendruscolo, 2008). Structural-based methods exploit structural constraints in the few solved three-dimensional conformation of amyloid fibrils (Thompson et al., 2006, Yoon and Welsh, 2004, Trovato et al., 2006, Galzitskaya et al., 2006a, Bryan et al., 2009).

The first mathematical tool was developed using *in vitro* experimental data of the model protein human muscle acylphosphatase (AcP) (Chiti et al., 2003). The differences observed on the aggregation rates of several point mutants were correlated with changes in their hydrophobicity, the propensity to form α-helical and β-sheet structure, and overall charge. The empirical formula obtained allows indentifying the propensity to form β-sheet aggregation in all unstructured or intrinsically disordered polypeptides. A further refinement of this equation included other extrinsic factors such as pH, ionic strength and protein concentration, allowing calculation of the rate constants for aggregation in different conditions (DuBay et al., 2004). Dubay's algorithm was used, considering only intrinsic properties, to determine the amyloid deposition tendency of the 20 natural amino acids, resulting in aggregation propensity profiles for the target proteins in which the aggregation-prone regions can be easily identified (Pawar et al., 2005). Zyggregator is a predictor method based on Pawar's algorithm that, combined with protein flexibility and solvent accessibility analyses from the CamP method (Tartaglia et al., 2007), allows identifying the aggregation tendency profile for a given protein in any conformational state (Tartaglia and Vendruscolo, 2008).

TANGO is another empirical server able to predict β-sheet aggregation by analyzing the propensity of each residue to acquire different structural elements, such as α-helix, β-turn, β-sheet aggregation and α-helical aggregation (Fernandez-Escamilla et al., 2004). This prediction method considers protein stability and intrinsic properties of polypeptide chains (hydrophobicity, electrostatic interactions, hydrogen bonds contributions, structural conformation propensities), as well as extrinsic factors (pH, ionic strength, peptide concentration). Combining this empirical information with structural studies of several amyloid-forming hexapeptides (AmylHex dataset), WALTZ algorithm was developed in order to distinguish between amorphous β-sheet aggregates and ordered amyloid-like deposits (Maurer-Stroh et al., 2010).

The SALSA algorithm was generated to locate regions with high β-strand propensity, assuming that formation of fibrillar aggregates is strongly associated to β-strand formation (Zibaee et al., 2007).

A large set of punctual mutants of amyloid β-peptide was used to calculate their relative solubility in the cytoplasm of *E. coli* (Sanchez de Groot et al., 2005). Thereafter,

the obtained scale of intracellular intrinsic aggregation propensity for the 20 natural amino acids allowed to develop the AGGRESCAN web server (Conchillo-Sole et al., 2007), in which an aggregation propensity profile is provided from the amino acid primary sequence. This software is the first predictive method totally based on *in vivo* empirical information, which permits to identify aggregation-prone regions of protein sequences *in vitro*.

The PASTA algorithm predicts the regions of polypeptide sequence involved in the formation of ordered cross-β structure (Trovato et al., 2006). A dataset of globular proteins with well-known secondary structures was used to calculate the pairing energies for each pair of residues facing one another on parallel or antiparallel neighbouring strands within a β-sheet. In contrast to most other methods, PASTA algorithm is able to identify whether β-strand in the fibrils is susceptible to adopt a parallel or antiparallel orientation.

Similar to the PASTA method, the BETASCAN algorithm was derived from a selected structures of amphipathic β-sheets in the PDB database and determines the preference for each pair of amino acids to be hydrogen bonded in a β-sheet, assuming that the sequence not only determines the secondary structure, but also the assembly of individual β-strand into ordered β-sheets (Bryan et al., 2009).

FOLDAMYLOID exploits a scale of packing density to identify aggregation-prone and intrinsically disordered regions in polypeptide primary sequences. As deduced from the analysis of the spatial structures of proteins in the PDB (Galzitskaya et al., 2006b), hydrophobic residues (Val, Met, Leu, Ileu, Tyr, Phe, and Trp) were found the most packed side chains in globular proteins.

The 3D profile method and ZIPPER database are based on the crystal structure analysis of the amyloid fibril forming hexapeptide NNQQNY, with the aim to predict regions of polypeptide sequences forming amyloid aggregates (Thompson et al., 2006). ZIPPER database consists of a large set of calculated predictions resulting from the analysis of 20000 different protein sequences using the 3D profile algorithm. The accuracy of the sequence predictions is evaluated in energetic term using the ROSETTADESIGN software (Kuhlman and Baker, 2000).

Overall, the information provided by these predictive programs has contributed significantly to understand the mechanism of protein deposition. The effect of natural

or designed mutations on individual proteins or large set of proteins may be easily studied, being thus of high interest for the biotechnological production of proteins and to develop therapeutic approaches for human amyloidoses.

6. - INTERACCION SURFACES

6.1 - Quaternary structure and disease

A large number of proteins contain more than one polypeptide chain, forming the quaternary structure by association of their subunits. The correctly association of polypeptides into functional complexes is an indispensable requisite to maintain homeostasis in living cells. By contrast, anomalous interactions can lead to protein misfolding and disease.

Negative selection as a mechanism for specificity is a very useful strategy considering the crowded environment inside cells. Several β-sheet proteins have used negative design principles to prevent aberrant assembly (Richardson and Richardson, 2002) and promote solubility (Doye et al., 2004). Thus, similar strategies, as an evolutionary mechanism, can be exploited to improve specificity in protein interactions (Baker, 2006).

Several amyloidogenic proteins present quaternary structure or are bound to other proteins under physiological conditions. More than 100 different SOD1 mutants are associated to FALS disease. A4V SOD1 variant has an increased tendency to aggregate, and interestingly, the mutation is close to the dimer interface promoting interface destabilization. This information supports the theory that only monomeric species are able to associate into amyloid fibrils, being an obligatory step for protein aggregation (Ray and Lansbury, 2004). Other amyloidogenic protein is the homotetrameric TTR, which causes familial amyloidotic polyneuropathy (FAP). Many of amyloidoses-related mutations favor destabilization of monomeric interfaces and promote complex dissociation.

6.2 - Interface structural properties and prediction algorithms

Protein-protein interaction sites have specific chemical and physical characteristics, all of which contribute to the molecular recognition process. Hydrogen bonds and van der Waals interactions contribute significantly to the free energy of protein-protein interactions, but the major factor in stabilizing protein interfaces is hydrophobicity (Chothia and Janin, 1975). Analysis of residue clustering on protein surfaces revealed that interface sites contain the most hydrophobic residue clusters of all those on the protein surface (Young et al., 1994). The interior, surface, and interface components in oligomeric proteins have been analyzed, specifically the contribution of structural properties such as hydrophobicity, accessible surface area, shape, and residue preferences have received attention (Janin et al., 1988, Miller, 1989, Argos, 1988). Size and shape of interfaces can be measured simply in absolute dimensions (Å) or in terms of accessible surface area (ΔASA). The concept of ASA describes the extent to which protein atoms can form contacts with the solvent, and is strongly correlated with hydrophobic free energies.

Predicting the position of the interface in dimeric proteins has been possible using a predictive algorithm based on the hydrophobicity of residue clusters in proteins (Young et al., 1994). Besides, electrostatic complementarity between interfaces has been employed as an additional filter for many protein-protein docking methods (Vakser and Aflalo, 1994), as well as new methods based on the evaluation of shape complementarity (Lawrence and Colman, 1993).

Overall, protein–protein interfaces are more hydrophobic, planar, globular and protruding than other parts of a protein's surface (Jones and Thornton, 1997a). This data led to develop a simple method for predicting protein–protein interactions based on six structural parameters: solvation potential; hydrophobicity; accessible surface area; residue interface propensity; planarity and protrusion (Jones and Thornton, 1997b).

The Optimal Docking Area (ODA) method identifies protein surface patches of different sizes (Fernandez-Recio et al., 2005). This algorithm is based on previously solvation studies derived from octanol/water transfer experiments and adjusted for protein-protein binding (Fernandez-Recio et al., 2004).

SHARP^2 is a web-based tool for predicting protein-protein interactions sites on protein structures (Murakami and Jones, 2006). The score for each patch of residues is a combination of six calculated parameters: solvation potential; hydrophobicity; accessible surface area; residue interface propensity; planarity and protrusion (SHARP^2).

Computational approaches such as alanine-scanning have shown that the amino acidic composition in interaction surfaces is enriched in Trp, Tyr and Arg residues (Bogan and Thorn, 1998). Besides, Trp, Phe and Met residues are highly conserved in protein interfaces, suggesting that hydrophobic forces are really important for the interaction surface of the monomers (Monsellier et al., 2008).

Protein-protein interfaces are stabilized mainly by hydrophobic and electrostatic interactions (Chothia and Janin, 1975, Jones and Thornton, 1996), which have been also indentified as the main forces modulating protein aggregation. Thus, some overlap between the aggregation-prone regions and the interaction surfaces might be expected since the nature of the residues that populate these regions appears to be largely similar.

7. - PROTEIN MODELS

7.1 - Leech carboxypeptidase inhibitor (LCI)

LCI is a powerful competitive inhibitor of different types of pancreatic-like carboxypeptidases: A1; A2; B and plasma carboxypeptidase B (pCPB). pCBP is also known as thrombin-activatable fibrinolysis inhibitor (TAFI), which leads to an attenuation of fibrinolysis. Consequently, this inhibitor may help to preserve blood in liquid state during leech nutrition.

This polypeptide is a 67-residues cysteine-rich protein which folds in a compact domain of five-stranded antiparallel β-sheet and a short α-helix. The major contribution for its stability is the presence of four disulfide bridges between cysteines 10-33, 17-61, 18-42, and 21-57, all of them connecting secondary structure elements (figure 11).

The oxidative folding pathway of LCI shares mechanistic similarities with both BPTI and hirudin proteins (Salamanca et al., 2003). Reduced and denatured LCI folds through a sequential flow of 1- and 2-disulfide intermediates, guiding the formation of two different population that act as kinetic traps: one is a mixture of four-disulfide (scrambled) species; and the other consists in three predominant three-disulfide isomers, in which two of them contain all-native disulfides. Like in the case of hirudin folding pathway, a heterogeneous mixture of 1- and 2- disulfide intermediates leads to the formation of 4-disulfide scrambled species, which might attain the native functional form. In contrast, and as in the case of BPTI, three native-disulfide intermediates populate the late steps of folding acting as major kinetic traps.

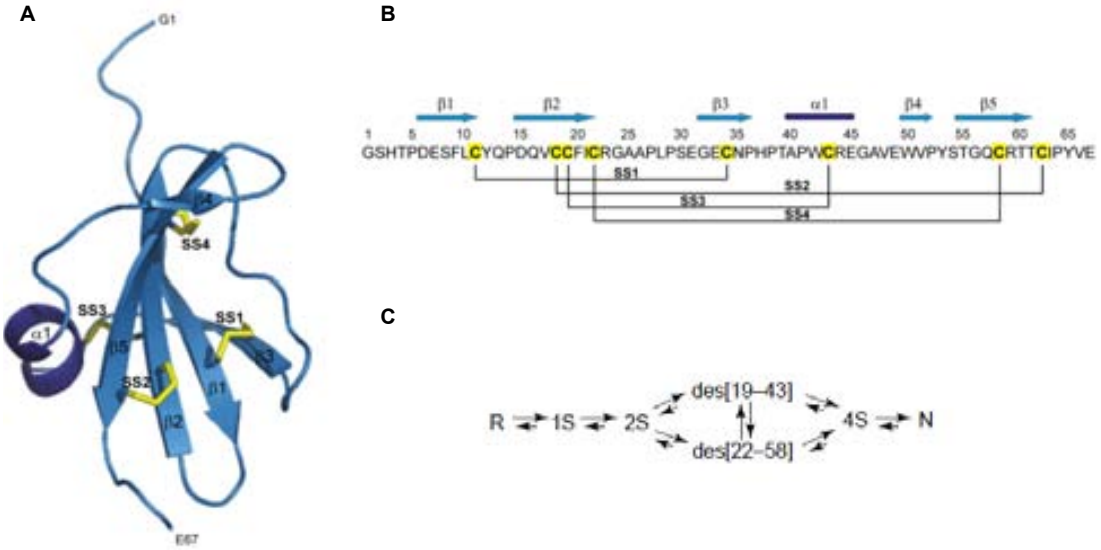


Figure 11. **A**. Ribbon illustration of the native three dimensional structure of LCI. (PDB: 1DTV). The yellow sticks represent the four native disulfide bonds. **B**. The secondary structure elements and disulfide pairing are shown above and below its amino acid sequence, respectively (Arolas et al., 2009).**C**. Disulfide folding pathway of LCI. R and N indicate the reduced and the native forms, respectively. 1S, 2S, 3S and 4S are ensemble of molecules with the corresponding number of disulfide bridges. 3S N are 3 native-disulfide species.

The two major kinetic traps (III-A and III-B) posses only native disulfides bonds and lack disulfides 22-58 and 19-43 in intermediates, respectively (Salamanca et al., 2003). Stop/go folding experiments demonstrate that the rate of interconversion between the two 3 native-disulfide intermediates are faster than the conversion into scrambled species (Arolas et al., 2004). Both intermediates are partially structured forms and

show high stability against denaturants. Their conformational properties were analyzed by CD, D/H exchange followed by matrix-assisted laser desorption/ionization time-of-flight mass spectroscopy (MALDI-TOF MS) and ^{1}H NMR spectroscopy (Arolas et al., 2004). By CD spectroscopy, the III-A intermediate shows a high percentage of residues in β-structure similar to that of the native conformation. On the contrary, the CD spectra of the III-B species and scrambled isomers present clear differences from that of the native form. The extent of protected deuterons in the native structure is higher than in both III-A and III-B forms, and scrambled isomers. The ^{1}H NMR spectra of native LCI and III-A intermediate display very similar signal dispersion, suggesting the properly folding of these species. However, III-B forms and scrambled species show peak collapse, indicating the presence of partially folded structures and random coil conformations. Overall, these conformational features have been recently supported by the structure resolution of the III-A intermediate by NMR (Arolas et al., 2005a), and the crystallographic resolution of the III-B analog intermediate (Arolas et al., 2005b).

7.2 - SH3 domain of α-spectrin protein (SPC-SH3)

The conformational stability and folding mechanism of the SH3 domain of α-spectrin have been extensively studied (Martinez et al., 1998, Sadqi et al., 1999, Periole et al., 2007). SPC-SH3 is a 62-residue polypeptide that folds into an orthogonal β-sandwich with a short 3_{10} helix (Musacchio et al., 1992). This domain has been largely characterized, especially its thermodynamic and kinetic features during the folding reaction, showing a two-state model with no significant accumulation of folding intermediates (Viguera et al., 1994). Its thermodynamic stability also depends strongly on the pH (Sadqi et al., 1999). The role of the hydrophobic core and the distal β-hairpin in its conformational stability has been carefully studied (Viguera et al., 1996, Grantcharova et al., 1998), describing the 3_{10} helix and the β-hairpin as the folding nucleus of the protein (Martinez and Serrano, 1999) (figure 12).

Protein engineering has been used to insert different poly-Gly length in the loop, revealing important issues about the high degree of compactness of the folding core (Martinez et al., 1999). Later on, the thermodynamic and kinetic stabilities of 20 *de novo* design mutants of the hydrophobic core were analyzed (Ventura et al., 2002). These variants contained single and multiple substitutions in the comprising 9 residues

of the core (Val9, Ala11, Val23, Met25, Leu31, Leu33, Val44, Val53 and Val58). The distal loop mutants (residues Ans47 and Asp48) displayed differential thermodynamic and conformational properties, as well as the core variants.

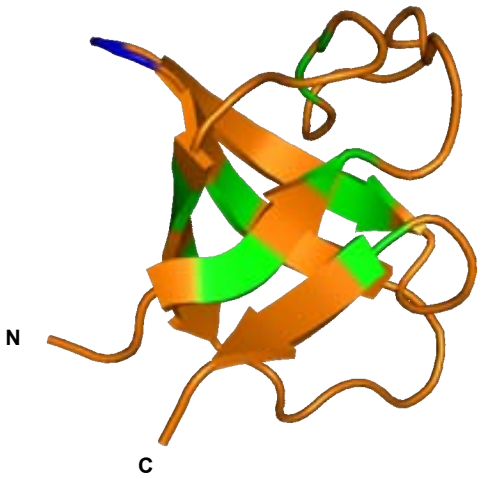


Figure 12. Ribbon diagram of the α-spectrin SH3 domain (PDB 1SHG). The green backbones correspond to the residues at the hydrophobic core. Residues in the distal β-hairpin are shown in blue. This figure has been prepared with PyMol.

7.3 - *FF domain of yeast URN1 protein (URN1-FF)*

During the synthesis and processing of mRNA several proteins regulate this process binding to the phosphorylated carboxyl-terminal domain (CTD) of eukaryotic RNA polymerase II, as the yeast Prp40 protein (Morris and Greenleaf, 2000) and the human CA150 protein (Goldstrohm et al., 2001). Several studies have demonstrated that these interactions are mediated by the presence of FF domains. These domains are present in several copies in other proteins containing WW domains, presenting two highly conserved phenylalanine residues and a characteristic three α-helix fold (α1-α2-3_{10}-α3) (Bedford and Leder, 1999).

Generally, FF domains act as protein-protein interaction modules present in three eukaryotic protein families: the splicing factors (human HYPA/FBP11, yeast Prp40 and yeast URN1); the transcription human factor CA150; and p190RhoGTPase-related proteins. They display large sequential divergence and involved in unrelated biological pathways since they can interact with different kind of ligands (phosphorylated CTD, TFII-I transcription factors and TPR motif among others).

The structures of several FF domains (from HYPA/FBP11, Prp40, URN1 and CA150) have been characterized in detail (Allen et al., 2002, Bonet et al., 2009, Bonet et al., 2008, Lu et al., 2009) (figure 13). The FF domain of the human HYPA/FBP11 protein has been extensively studied, in particular its transition state for folding using fast-flow denaturation methods and molecular dynamics (Jemth et al., 2004, Jemth et al., 2005, Korzhnev et al., 2007, Jemth et al., 2008, Korzhnev et al.).

The URN1 protein is a pre-mRNA splicing factor from *Saccharomyces cerevisiae*, which contains only two protein domains: a WW and a FF domain. A large number of proteins interact with the URN1 protein, for instance, it binds directly to the spliceosomal subcomplex Prp19 protein (Ren et al.). However, it has not be described any specific partner for the URN1-FF domain. This 59-residue protein shows the typical FF fold α1-α2-3_{10}-α3 but differs from other FF domains on its charge distribution in the surface (Bonet et al., 2008).

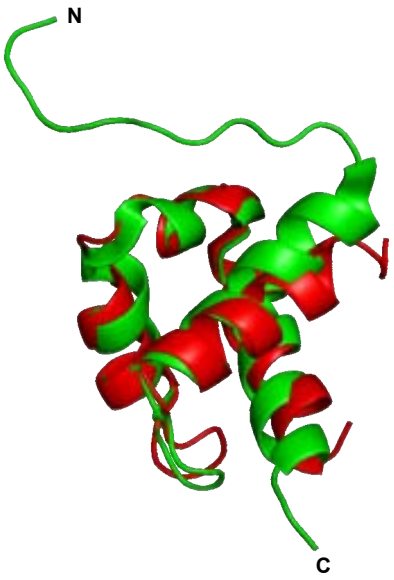


Figure 13. Ribbon diagram of two superimposed structures: the FF domain of the yeast URN1 and the human HYPA/FBP11 proteins; in red and green respectively. The PDB accession code for URN1-FF structure is 2JUC, and for FBP11-FF is 1UZC. This figure has been prepared with PyMol.

II - AIMS

The global aim of this thesis is to provide further insights about the physicochemical principles controlling the balance between the population of native soluble conformations and toxic aggregated species in globular proteins under physiological conditions.

More specifically, the following objectives were proposed for each scientific paper:

Paper I:

- To study the role of the individual disulfide bridges in the folding and the stability of the LCI protein.

Paper II and IV:

- To correlate the thermodynamic stability of a set of SPC-SH3 variants with their in vivo deposition and their in vitro aggregation levels.
- To analyze the structural properties of in vivo SPC-SH3 aggregates.
- To correlate the aggregation inside bacteria of SPC-SH3 domains with their kinetic folding properties.

Paper III:

- To analyze the spatial coincidence between aggregation-prone regions and interaction surfaces in multimeric proteins associated with conformational diseases.

Paper V:

- To study the conformational conversion of native α-helices into amyloid β-sheet structures using the all-α FF domain of URN1 as a protein model.
- To identify the region controlling the balance between solubility and aggregation in this domain.

III - DISCUSSION

Paper I: The role of disulfide bridges on the LCI conformational stability

In this thesis, the role of disulfide bridges during the acquisition of the native structure, conformational stability and function have been characterized using LCI as a model, a small four-disulfide polypeptide. Different disulfide delection mutants have been analyzed using oxidative and reductive unfolding approaches, disulfide scrambling techniques and activity assays.

The overall data reveals that not all the covalent bridges are equally important for the stability of the LCI protein. The elimination of specific disulfide bridges results in a strong structural destabilization. The protein is able to attain a native-like conformation and function when the individual disulfide bonds SS3 and SS4 are missing, which interestingly correspond to those reduced in III-B and III-A folding intermediates, respectively. Elimination of the disulfide bond SS1 has the strongest effect on the conformational stability, being the corresponding mutant unable to gain the native functional structure. The LCI mutant lacking the SS2 does not accumulate during the oxidative folding reaction, despite being conformationally stable.

Disulfide-containing proteins functioning as carboxypeptidase inhibitors are designed to be highly stable under extreme external conditions. LCI protein possesses a high stability caused by a compact hydrophobic core and an accurate disulfide-bond distribution (Salamanca et al., 2002). Interestingly, it is shown that the stability and biological function of LCI protein are not necessary coupled. In this sense, single LCI mutants lacking SS2 and SS3 disulfides show excellent inhibitory activity, despite their significant destabilization.

Reduction of disulfides buried in the folded structure uses to be the rate-limiting step during the unfolding reaction of disulfide containing proteins. Resistance to reduction was found to agree with the changes in conformational stability of single LCI mutants, the lower the conformational stability of the mutant, the higher the sensitivity of its disulfides to the reducing agent.

These experimentally results were compared with the predicted stabilities of mutant LCI variants upon removal of one or more disulfides using the Wang-Uhlenbeck formula for loop entropy (Wang and Uhlenbeck, 1945). The predicted changes in stability for LCI mutants were in contradiction with experimental data, indicating that

other destabilizing contributions, apart from the gain in entropy in the unfolded state should account for these differences. The contribution of other forces and interactions to conformational stability were addressed using the FoldX forcefield (Guerois et al., 2002, Schymkowitz et al., 2005). This algorithm accurately predicts the rank and relative differences in stability for those LCI mutants able to fold into a predominant conformation.

Our results also provide information about the role of disulfide bonds in the folding pathway of LCI. Previous studies revealed the accumulation of two major intermediates along the folding reaction, the III-A and III-B species, which have the ability to protect their three native disulfides from rearrangement (Salamanca et al., 2003, Arolas et al., 2004). Surprisingly, single mutants lacking SS1 and SS2 disulfide bridges do not accumulate during the folding pathway, indicating that conformational stability is not enough to attain the native disulfide connectivity. Regarding the formation mechanism of III-A and III-B intermediates, it has been demonstrated that 3S species with non-native disulfides can rearrange to form 3S natively bonded species, suggesting that the non-specific oxidation of 2S forms into 3S species and the further reshuffling of this ensemble could be the underlying mechanism accounting for the formation of both intermediates.

Paper II and IV: *In vivo* and *in vitro* SPC-SH3 aggregation correlates with the thermodynamic and kinetic stabilities

In the first part of this section we have used a set of 23 SPC-SH3 variants to analyze the determinants of protein aggregation under physiological conditions. These mutants showed differential *in vivo* solubility, which allowed classifying them into three sets: a) variants with similar distribution between the soluble and insoluble fractions, including SPC-SH3 *wt*; b) mutants predominantly or totally present in the insoluble fraction; and c) variants that mainly remained soluble.

Efforts were made to correlate the aggregation propensity with the physicochemical parameters of the primary sequence. As an example, global hydrophobicity and the aliphatic index were not found to be associated with the

solubility of the variants. In a similar way, the aggregation propensities were predicted using AGGRESCAN and TANGO algorithms, but no relationship existed between the predicted tendency to aggregate and the fractioning of the variants in the cellular lysates. These computational tools assume that aggregation is initiated by association of aggregation-prone regions exposed to the solvent, thus suggesting that SPC-SH3 variants might aggregate from partially unfolded species.

A tightly correlation between *in vivo* aggregation and thermodynamic stability was observed. Indeed, destabilization of globular proteins is expected to totally or partially unfold the native states, being the conformational stability a controlling factor for *in vivo* protein aggregation. This result is in good agreement with these obtained using other protein models, such as HypF-N and p53 proteins (Calloni et al., 2005, Mayer et al., 2007). *In vitro* amyloid aggregates were obtained under mild denaturing conditions, and their conformational properties were analyzed using far-UV CD, FTIR, TEM and binding to amyloid specific dyes. A striking association was observed between *in vitro* and in the cell aggregation properties, indicating that both processes are modulated by the conformational stability. Interestingly enough, these self-assembly reactions do not require a large destabilization that completely unfold the globular domains, indicating that subtle changes in thermodynamic stability can induce local fluctuations displacing the equilibrium towards aggregated states.

In the paper III, our purpose was to decipher whether the energy barriers to unfolding are equally or more important than the thermodynamic stability for the aggregation of small globular proteins inside the cell, since conformational and kinetic stability are usually correlated. Indeed, aggregation promoting mutations of a reduced number of globular proteins have been related with changes in the unfolding barriers but not in the thermodynamic stability (Foss et al., 2005, Liemann and Glockshuber, 1999).

We completed previous studies (Ventura et al., 2002) by characterizing the folding kinetics of destabilized SPC-SH3 variants to collect a complete set of kinetic data corresponding to both stabilized and destabilized variants. This allowed us to analyze the contribution of the kinetic parameters to the intracellular aggregation of this globular protein. As described above, a good correlation was observed between the formation of intracellular deposits and their thermodynamic stability. Assuming that

aggregation occurs during the folding reaction, it should be expected a good correlation between the folding activation free energy and the experimental solubility. However a lower correlation was observed in this case. Regarding the association between the unfolding kinetic barriers and the *in vivo* solubility, the observed correlation was better than that found for the refolding kinetic barriers, but still lower than that obtained when considering thermodynamic stability alone. By analyzing the intracellular deposition of a SPC-SH3 mutant displaying a random coil structure (Eichmann et al., 2010), it has been demonstrated that the aggregation of this globular protein does not requires the population of molten-globule like intermediates. The Lumry-Eyring model accurately describes the observed behavior (Sanchez-Ruiz, 1992) implying that the aggregation rate is very slow relative to the unfolding rates of the mutants. As a result, protein aggregation becomes almost independent of the unfolding and refolding rates and depends only on the conformational stability.

We also studied the effect of temperature on the solubility of SPC-SH3 variants. The soluble cell fractions of several variants were incubated at different temperatures and the aggregation kinetics were monitored. The observed aggregational behaviors were in agreement with their relative thermal denaturation stabilities.

Finally, the thermodynamic stabilities of purified SPC-SH3 variants and their cellular lysates were studied using pulse proteolysis. Two different experiments were performed: one using different concentrations of denaturant; and the other one following the kinetics of proteolysis in the absence of denaturant. The conformational stabilities against chemical denaturation and under physiological-like conditions appeared to be largely similar, highlighting the existing association between *in vitro* and *in vivo* protein conformational properties.

Overall, these results propose thermodynamic stability as the main controlling factor during intracellular protein deposition, indicating that, for small proteins, the energy barriers for unfolding should be more important than the refolding ones, since usually these molecules emerge from the ribosome as folded and globular species.

Paper III: Overlapping aggregation regions and interaction surfaces in disease-related proteins

In this work we have measured the spatial coincidence between high-propensity aggregation regions and the predicted and real protein interfaces in globular proteins associated with conformational diseases.

Computational methods have become powerful tools in predicting aggregation-prone regions in globular proteins, as well as detecting interaction regions in protein surfaces. Here, we have used TANGO and AGGRESCAN web servers, and the algorithms implemented by Galzitskaya et al. and Zhang et al. All of them allow predicting, from the primary sequence, the regions partially or totally exposed to solvent which are able to nucleate the self-assembly reaction. Regarding to the prediction of protein-protein interaction interfaces, three structure-based methods were used: ODA; SHARP2; and InterProSurf. In order to evaluate the degree of proximity between an aggregation-prone region and the interface, we defined the Interface Proximity Index (IPI).

Protein-protein interaction surfaces and regions with high propensity to aggregate have been analyzed in several amyloidoses-related proteins. β2m protein can form an extracellular ternary complex with the HLA heavy chain, or inside the cell β2-m can be associated with the HFE protein. In both cases, more than 75% of the residues comprising the two aggregation-prone regions are less than 3 Å from the interface of the complexes. In addition, the high IPI values of the regions with high tendency to aggregate reflect that they are preferentially located close to the interaction sites. The native homotetramer transthyretin encloses five predicted aggregation-prone regions, in which, with one exception, 90% of the residues in the aggregating patches are at less than 3 Å from the interactions regions. SOD1 protein is a homodimer displaying four aggregating *hot spots*, in which the 61% of the encompassing residues are close to the interfaces. All of the predicted aggregating regions are highly exposed to the solvent in the monomeric form, and except one of them, they show high IPIs. Other interesting proteins are the light chains (LCs) of immunoglobulins and the human IgG1 antibody, which are associated to AL amyloidosis (Sanchorawala, 2006, Eulitz et al., 1990) and a systemic amyloid disease, respectively. Free secreted LCs molecules can form homodimers which in certain occasions can aggregate and become cytotoxic.

Two of the five detected aggregation-prone regions in the IgG1 light chain are located close to the dimer interface. In the case of the IgG1 heavy chain, four of the nine aggregating regions show high IPI values, and interestingly, they are located in the heavy-chain variable domain (VH) close to the heterotetramer interface.

We show that succesfull strategies to avoid protein aggregation and disease have in fact exploited the generation of new binding interfaces. A camelid antibody fragment has been demonstrated to inhibit the *in vitro* aggregation of a disease-related lysozyme variant (Dumoulin et al., 2003). Interestingly, the interaction region is not located near the site of the mutation neither in the destabilized protein region, but the antibody prevents structural fluctuations and exposure of amyloidogenic regions. Similarly, the two aggregation-prone regions of Aβ peptide can be protected by using the Z domain derived from protein A (affibody) (Hoyer et al., 2008), which allows burying the aggregation *hot spots* within a large hydrophobic cavity.

In order to decipher whether the found overlapping between aggregation-prone regions and protein-protein interaction sites is restricted only to amyloidogenic proteins, we analyzed monomeric and globular proteins which are non-related with conformational diseases. In cases such as human myoglobin and maltose binding protein (MBP), predicted interfaces and aggregation regions do not overlap, and most of the aggregating residues are buried in the hydrophobic core. We also show that, evolution has used negative design to avoid protein deposition when aggregation prone regions should be exposed to solvent in order to facilitate binding to specific protein targets. This strategy can be found in thioredoxin A and ubiquitin proteins. Finally, the analysis of 25 non-amyloidogenic dimeric proteins showed that almost all of the homodimers have at least one aggregation region in which more than 85% of the residues are close to the interaction surfaces.

Overall, different amyloidoses-related proteins have been studied revealing a significant overlapping between aggregation *hot spots* and interactions regions. These evidences suggest that the formation of native complexes might compete with aggregation under physiological conditions. Thus, stabilizing protein interfaces could be a good strategy to avoid protein deposition and cellular toxicity.

Paper V: Transition between the soluble and amyloid states of a small all-α domain

The transition between all-α and amyloid structures underlies the onset of several human neurodegenerative disorders. In the present work, we have used the all-α FF domain of yeast URN1 as a model system and a battery of biophysical techniques to analyze the conformational conversion of native α-helices into amyloid β-sheets, providing new data on the structural features of early species in amyloid formation processes.

Native and globular proteins usually require a partial or total unfolding to form amyloid fibrils, this effect can be promoted either by mutations or by external factors. Here, we have analyzed the aggregation propensity of the URN1-FF protein over at acidic pH, observing that at physiological temperature the formation of β-sheet enriched amyloid fibrils only occurs below pH 3.0.

The conformational analysis of soluble URN1-FF species at different pHs and low concentration indicate that, above pH 2.0, the soluble population maintains a significant amount of its native α-helical content without detection of the formation of β-sheet structures for 24h, revealing that they are at least metastable. At pH 3.0 and 2.5, the protein keeps essentially the same α-helical content than the native protein, however thermodynamic data indicate that the protein is destabilized in these conditions. At pH 2.0, the α-helical content decreases slightly and the protein stability is significantly reduced. Chemical denaturation experiments demonstrate that, at 37 °C, ~50% and ~10% of URN1-FF species are already unfolded at pH 2.0 and pH 2.5, respectively. Overall, the species populated at low pH correspond to molten globule-like conformations.

Several computational algorithms detected a main aggregation-prone region inside the helix 1 of URN1-FF. This is also the region with the highest predicted helical propensity. Thus, this domain does not exhibit the kinetic partitioning of protein folding and aggregation reactions characteristic of other amyloidogenic proteins (Chiti et al., 2002). The results illustrate how nature finds difficult to avoid the presence of amyloidogenic stretches in proteins because, at least in certain cases, the regions leading the formation of native structures are also able to trigger self-assembly into

toxic conformers. As illustrated here, in these cases it is the stability of the native state that prevents the transition towards amyloidogenic conformations in physiological conditions. The data suggest that the stabilization of α-helical structures might become a useful approach to modulate the aggregation of disease-linked polypeptides.

IV - GENERAL DISCUSSION

During the last decades, protein folding and aggregation have become important research areas in biomedicine since a huge number of fatal diseases are linked to protein deposition processes. Moreover, the formation of inclusion bodies during protein recombinant expression has become a major bottleneck in the production of protein-based drugs by the biotechnological industry; although new evidence suggest some than these *in vivo* aggregates might be useful for different purposes. In the present thesis, we have focused our efforts in understanding how different model proteins fold and aggregate under physiological conditions, as well as in deciphering the physicochemical properties that control these reactions. Towards this aim, we have used *Escherichia coli* as a simple but physiologically relevant *in vivo* experimental model as well as a combination of biochemical, biophysical and computational approaches.

The study of the role of disulfide bridges during the acquisition of the native structure of a small disulfide-rich protein has revealed that not all the disulfides contribute equally to the global stability of the functional native protein. The presence of disulfide bonds on protein structures is supposed to increase the thermodynamic stability by decreasing the conformational entropy of the denatured state. Using the LCI protein as a model, we have demonstrated that the elimination of disulfide bonds can lead to dramatic effects on the conformational stability of disulfide proteins, but other factors apart from entropic contributions should be taken into account to rationalize their effect.

Our analysis highlights the importance of thermodynamic stability on protein solubility under physiological conditions. The negative correlation observed between protein stability and aggregation has significant interest in the fields of protein production and therapeutics for amyloid diseases. Moreover, we have demonstrated that, at least in some globular proteins, the *in vivo* and *in vitro* solubilities are modulated by the same physicochemical properties. The data suggest that stabilization of the native state of globular proteins, preventing local and global fluctuations that expose aggregation-prone regions to the solvent, might become a promising strategy for therapeutic intervention in conformational disorders.

This thesis has provided support to the idea that interaction interfaces are more aggregation prone than other surface regions in proteins complexes. It is suggested

that, specially for homo-multimeric proteins, after protein synthesis and folding, monomeric species should associate rapidly into the native complexes, to avoid the prolonged exposure of aggregation-prone regions to the solvent. Accordingly, we propose that the formation of protein complexes can be considered as a protective strategy against protein aggregation inside the cell. Many of amyloidoses-related mutations are associated to destabilization of protein-protein interaction surfaces, suggesting the over-stabilization of interfaces being a good approach to compete protein misfolding and disease.

The present thesis has also addressed the conversion of an all-α protein into amyloid β-sheet structures. Evolutionarily, it has been difficult to avoid the presence of regions with high tendency to aggregate in proteins, likely because the non-covalent contacts that stabilize native structures resemble those leading to the formation of amyloids. We show here that, in certain cases, diverging structural properties like native helical conformation and amyloid β-sheets propensity might overlap in the same protein region. Hence again, the conformational stability appears as an essential factor to prevent the formation of aberrant aggregates, since by favouring the population of the native conformation at equilibrium it prevents significant exposure of sticky regions able to establish the initial β-sheets intermolecular contacts leading to aggregation.

V - CONCLUDING REMARKS

Paper I: The role of disulfide bridges on the LCI conformational stability

- Using oxidative and reductive unfolding approaches we have demonstrated that not all the bridges are equally important for the stability of a small disulfide-rich protein.
- Generally, the elimination of individual disulfide bridges results in a strong structural destabilization, as in the mutant lacking the SS1 disulfide which is not able to attain the native structure. However, native-like conformations can be obtained when individual disulfide bonds SS3 and SS4 are missing, which interestingly correspond to those absent in III-B and III-A folding intermediates, respectively.
- The predicted gain in entropy in the unfolded state does not suffices to explain the thermodynamic data, indicating that other contributions apart from entropy, such as van der Waals interactions, might influence the global stability of LCI mutants.
- As a general trend, the accumulation of stable intermediates during the LCI folding reaction relies on their ability to protect the native disulfides from further rearrangement. However, the LCI_2 mutant does not accumulate during the folding reaction, despite being conformationally stable. This suggests that, in certain cases, the achievement of native disulfide connectivity might be regulated by kinetic instead of thermodynamic constraints.
- Surprisingly, LCI_2 and LCI_3 possess good inhibitory activity despite being significantly destabilized, indicating that the conformational stability and the inhibitory activity are not linked in this highly stable carboxypeptidase inhibitor.
- Previously, it was proposed that the formation of the two major LCI folding intermediates were formed by direct oxidation of natively-bonded 2S species. The present work proposes a modified LCI folding pathway: III-A and III-B intermediates arise from a non-specific oxidation of the 2S ensemble to non-natively bonded 3S species and the subsequent reshuffling into 3S species containing native disulfides.

Paper II and IV: *In vivo* and *in vitro* SPC-SH3 aggregation correlates with the thermodynamic and kinetic stabilities

Paper II:

- A set of 23 SPC-SH3 variants showing differential in vivo solubility has been classified into three groups: a) mutants with similar distribution between the soluble and insoluble fractions; b) variants that principally remained in the insoluble fraction; and c) variants that mainly remained soluble.
- The experimental solubility of the mutants cannot be correlated with the physicochemical parameters of the primary sequence neither the predicted aggregation propensities, suggesting that the deposition of the SPC-SH3 variants might be initiated from partially unfolded species.
- The thermodynamic stability of the variants clearly correlated with their *in vivo* deposition. Interestingly, a tight association was observed between the distribution of the protein in the soluble/insoluble fractions in cellular lysates and their *in vitro* aggregation, indicating that both phenomena are modulated by the conformational stability.
- Overall, this work demonstrates how the stability of native states prevents the unfolding and aggregation of globular proteins. A good strategy to delay the progress of depositional diseases might be the stabilization of native globular proteins.

Paper IV:

- *In vivo* protein aggregation correlates best with unfolding kinetic barriers than with folding activation free energies.
- We have used a SPC-SH3 mutant displaying random coil structure to demonstrate that, in this case, deposition occurs from essentially unfolded species, suggesting that the global aggregation rates can not be faster than the ones for unfolding.

- A good strategy to reduce or avoid *in vivo* protein deposition in small globular proteins would consist in lowering the cell growth temperature, since the aggregation rates are positively correlated with temperature.
- Pulse proteolysis indicates that the conformational and aggregational properties inside the cell and *in vitro* are similar, validating thus *in vitro* studies to approximate *in vivo* conditions.
- Overall, it can be concluded that the main factor controlling protein deposition in this model system is the thermodynamic stability at equilibrium.

Paper III: Overlapping aggregation regions and interaction surfaces in disease-related proteins

- Using computational methods, a stricking spatial coincidence has been found between the aggregation-prone regions and the interaction surfaces in the quaternary structure of amyloidoses-related proteins.
- The overlap between amyloidogenic regions and protein-protein interaction sites was also observed in proteins not related to disease.
- These results suggest that the formation of native complexes might compete with the deposition of the monomeric species, proposing that the acquisition of the native conformation in oligomeric proteins might be a protective strategy against aggregation.
- Several depositional diseases are related to mutations which affect the interface or the stability of protein complexes. Hence, the stabilization of protein interfaces appears as a promising approach to prevent the conversion of globular proteins into toxic assemblies.

Paper V: Transition between the soluble and amyloid states of a small all-α domain

- The stability of the URN1-FF soluble species is strongly affected by the pH, as observed by thermal and chemical denaturation assays. The species present

under acidic conditions exhibit essentially molten-globule like and partially unfolded conformations.

- Mild denaturation at low pH and physiological temperature induces the formation of URN1-FF ordered amyloid fibrils.
- The exposure of hydrophobic patches to solvent in globule-like conformations does not suffice to enable the conversion from native α-helices to β-sheet structures. This conversion needs at least the presence of partially unfolded conformations. However the formation of β-sheet amyloid aggregates does not require the population of a large amount of unfolded species at equilibrium.
- Limited proteolysis experiments revealed significant conformational flexibility in helix 1 and helix 3 of URN1-FF at low pH and computational approaches suggested helix 1 being the main aggregation-prone region, as well as the segment with the highest α-helical propensity.
- The dissection of this domain into its secondary structural elements has demonstrated that helix 1 plays a major role in controlling the conformation and solubility of this domain.

VI - REFERENCES

Allen, M., Friedler, A., Schon, O. & Bycroft, M. (2002). The structure of an FF domain from human HYPA/FBP11. *J Mol Biol, 323*, 411-416.

Anfinsen, C. B., Haber, E., Sela, M. & White, F. H., Jr. (1961). The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc Natl Acad Sci U S A, 47*, 1309-1314.

Argos, P. (1988). An investigation of protein subunit and domain interfaces. *Protein Eng, 2*, 101-113.

Arie, J. P., Miot, M., Sassoon, N. & Betton, J. M. (2006). Formation of active inclusion bodies in the periplasm of Escherichia coli. *Mol Microbiol, 62*, 427-437.

Arolas, J. L., Aviles, F. X., Chang, J. Y. & Ventura, S. (2006a). Folding of small disulfide-rich proteins: clarifying the puzzle. *Trends Biochem Sci, 31*, 292-301.

Arolas, J. L., Bronsoms, S., Aviles, F. X., Ventura, S. & Sommerhoff, C. P. (2008). Oxidative folding of leech-derived tryptase inhibitor via native disulfide-bonded intermediates. *Antioxid Redox Signal, 10*, 77-85.

Arolas, J. L., Bronsoms, S., Lorenzo, J., Aviles, F. X., Chang, J. Y. & Ventura, S. (2004). Role of kinetic intermediates in the folding of leech carboxypeptidase inhibitor. *J Biol Chem, 279*, 37261-37270.

Arolas, J. L., Bronsoms, S., Ventura, S., Aviles, F. X. & Calvete, J. J. (2006b). Characterizing the tick carboxypeptidase inhibitor: molecular basis for its two-domain nature. *J Biol Chem, 281*, 22906-22916.

Arolas, J. L., Castillo, V., Bronsoms, S., Aviles, F. X. & Ventura, S. (2009). Designing out disulfide bonds of leech carboxypeptidase inhibitor: implications for its folding, stability and function. *J Mol Biol, 392*, 529-546.

Arolas, J. L., D'Silva, L., Popowicz, G. M., Aviles, F. X., Holak, T. A. & Ventura, S. (2005a). NMR structural characterization and computational predictions of the major intermediate in oxidative folding of leech carboxypeptidase inhibitor. *Structure, 13*, 1193-1202.

Arolas, J. L., Popowicz, G. M., Bronsoms, S., Aviles, F. X., Huber, R., Holak, T. A. & Ventura, S. (2005b). Study of a major intermediate in the oxidative folding of leech carboxypeptidase inhibitor: contribution of the fourth disulfide bond. *J Mol Biol, 352*, 961-975.

Baker, D. (2006). Prediction and design of macromolecular structures and interactions. *Philos Trans R Soc Lond B Biol Sci, 361*, 459-463.

Baldwin, R. L. (1989). How does protein folding get started? *Trends Biochem Sci, 14*, 291-294.

Banci, L., Bertini, I., D'Amelio, N., Gaggelli, E., Libralesso, E., Matecko, I., Turano, P. & Valentine, J. S. (2005). Fully metallated S134N Cu,Zn-superoxide dismutase displays abnormal mobility and intermolecular contacts in solution. *J Biol Chem, 280*, 35815-35821.

Barnhart, M. M. & Chapman, M. R. (2006). Curli biogenesis and function. *Annu Rev Microbiol, 60*, 131-147.

Bedford, M. T. & Leder, P. (1999). The FF domain: a novel motif that often accompanies WW domains. *Trends Biochem Sci, 24*, 264-265.

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

Bemporad, F. & Chiti, F. (2009). "Native-like aggregation" of the acylphosphatase from Sulfolobus solfataricus and its biological implications. *FEBS Lett, 583*, 2630-2638.

Bemporad, F., Vannocci, T., Varela, L., Azuaga, A. I. & Chiti, F. (2008). A model for the aggregation of the acylphosphatase from Sulfolobus solfataricus in its native-like state. *Biochim Biophys Acta, 1784*, 1986-1996.

Bogan, A. A. & Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces. *J Mol Biol, 280*, 1-9.

Bonet, R., Ramirez-Espain, X. & Macias, M. J. (2008). Solution structure of the yeast URN1 splicing factor FF domain: comparative analysis of charge distributions in FF domain structures-FFs and SURPs, two domains with a similar fold. *Proteins, 73*, 1001-1009.

Bonet, R., Ruiz, L., Morales, B. & Macias, M. J. (2009). Solution structure of the fourth FF domain of yeast Prp40 splicing factor. *Proteins, 77*, 1000-1003.

Bryan, A. W., Jr., Menke, M., Cowen, L. J., Lindquist, S. L. & Berger, B. (2009). BETASCAN: probable beta-amyloids identified by pairwise probabilistic analysis. *PLoS Comput Biol, 5*, e1000333.

Calloni, G., Zoffoli, S., Stefani, M., Dobson, C. M. & Chiti, F. (2005). Investigating the effects of mutations on protein aggregation in the cell. *J Biol Chem, 280*, 10607-10613.

Canet, D., Last, A. M., Tito, P., Sunde, M., Spencer, A., Archer, D. B., Redfield, C., Robinson, C. V. & Dobson, C. M. (2002). Local cooperativity in the unfolding of an amyloidogenic variant of human lysozyme. *Nat Struct Biol, 9*, 308-315.

Carrio, M. M., Corchero, J. L. & Villaverde, A. (1998). Dynamics of in vivo protein aggregation: building inclusion bodies in recombinant bacteria. *FEMS Microbiol Lett, 169*, 9-15.

Carrio, M. M., Cubarsi, R. & Villaverde, A. (2000). Fine architecture of bacterial inclusion bodies. *FEBS Lett, 471*, 7-11.

Carrio, M. M. & Villaverde, A. (2002). Construction and deconstruction of bacterial inclusion bodies. *J Biotechnol, 96*, 3-12.

Carrio, M. M. & Villaverde, A. (2005). Localization of chaperones DnaK and GroEL in bacterial inclusion bodies. *J Bacteriol, 187*, 3599-3601.

Castillo, V., Grana-Montes, R., Sabate, R. & Ventura, S. (2011). Prediction of the aggregation propensity of proteins from the primary sequence: aggregation properties of proteomes. *Biotechnol J, 6*, 674-685.

Claessen, D., Rink, R., de Jong, W., Siebring, J., de Vreugd, P., Boersma, F. G., Dijkhuizen, L. & Wosten, H. A. (2003). A novel class of secreted hydrophobic

proteins is involved in aerial hyphae formation in Streptomyces coelicolor by forming amyloid-like fibrils. *Genes Dev, 17*, 1714-1726.

Colon, W. & Kelly, J. W. (1992). Partial denaturation of transthyretin is sufficient for amyloid fibril formation in vitro. *Biochemistry, 31*, 8654-8660.

Conchillo-Sole, O., de Groot, N. S., Aviles, F. X., Vendrell, J., Daura, X. & Ventura, S. (2007). AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics, 8*, 65.

Corazza, A., Rosano, C., Pagano, K., Alverdi, V., Esposito, G., Capanni, C., Bemporad, F., Plakoutsi, G., Stefani, M., Chiti, F., Zuccotti, S., Bolognesi, M. & Viglino, P. (2006). Structure, conformational stability, and enzymatic properties of acylphosphatase from the hyperthermophile Sulfolobus solfataricus. *Proteins, 62*, 64-79.

Coustou, V., Deleu, C., Saupe, S. & Begueret, J. (1997). The protein product of the het-s heterokaryon incompatibility gene of the fungus Podospora anserina behaves as a prion analog. *Proc Natl Acad Sci U S A, 94*, 9773-9778.

Creighton, T. E. (1990). Protein folding. *Biochem J, 270*, 1-16.

Creighton, T. E. & Goldenberg, D. P. (1984). Kinetic role of a meta-stable native-like two-disulphide species in the folding transition of bovine pancreatic trypsin inhibitor. *J Mol Biol, 179*, 497-526.

Chang, J. Y. (1993). Identification of productive folding intermediates which account for the flow of protein folding pathway. *J Biol Chem, 268*, 4043-4049.

Chang, J. Y. (1994). Controlling the speed of hirudin folding. *Biochem J, 300 (Pt 3)*, 643-650.

Chang, J. Y. (1996). The disulfide folding pathway of tick anticoagulant peptide (TAP), a Kunitz-type inhibitor structurally homologous to BPTI. *Biochemistry, 35*, 11702-11709.

Chang, J. Y. (2004). Evidence for the underlying cause of diversity of the disulfide folding pathway. *Biochemistry, 43*, 4522-4529.

Chang, J. Y. (2008). Diversity of folding pathways and folding models of disulfide proteins. *Antioxid Redox Signal, 10*, 171-177.

Chang, J. Y. & Li, L. (2005). Divergent folding pathways of two homologous proteins, BPTI and tick anticoagulant peptide: compartmentalization of folding intermediates and identification of kinetic traps. *Arch Biochem Biophys, 437*, 85-95.

Chang, J. Y., Li, L. & Lai, P. H. (2001). A major kinetic trap for the oxidative folding of human epidermal growth factor. *J Biol Chem, 276*, 4845-4852.

Chang, J. Y., Ventura, S. (2011). Folding of disulfide proteins. *Protein Reviews, 14*.

Chapman, M. R., Robinson, L. S., Pinkner, J. S., Roth, R., Heuser, J., Hammar, M., Normark, S. & Hultgren, S. J. (2002). Role of Escherichia coli curli operons in directing amyloid fiber formation. *Science, 295*, 851-855.

Chatrenet, B. & Chang, J. Y. (1992). The folding of hirudin adopts a mechanism of trial and error. *J Biol Chem, 267*, 3038-3043.

Chatrenet, B. & Chang, J. Y. (1993). The disulfide folding pathway of hirudin elucidated by stop/go folding experiments. *J Biol Chem, 268*, 20988-20996.

Chien, P., Weissman, J. S. & DePace, A. H. (2004). Emerging principles of conformation-based prion inheritance. *Annu Rev Biochem, 73*, 617-656.

Chiti, F. & Dobson, C. M. (2006). Protein misfolding, functional amyloid, and human disease. *Annu Rev Biochem, 75*, 333-366.

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

Chiti, F., Mangione, P., Andreola, A., Giorgetti, S., Stefani, M., Dobson, C. M., Bellotti, V. & Taddei, N. (2001). Detection of two partially structured species in the folding process of the amyloidogenic protein beta 2-microglobulin. *J Mol Biol, 307*, 379-391.

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.

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

Chiti, F., Taddei, N., Bucciantini, M., White, P., Ramponi, G. & Dobson, C. M. (2000). Mutational analysis of the propensity for amyloid formation by a globular protein. *EMBO J, 19*, 1441-1449.

Chiti, F., Webster, P., Taddei, N., Clark, A., Stefani, M., Ramponi, G. & Dobson, C. M. (1999). Designing conditions for in vitro formation of amyloid protofilaments and fibrils. *Proc Natl Acad Sci U S A, 96*, 3590-3594.

Chothia, C. & Janin, J. (1975). Principles of protein-protein recognition. *Nature, 256*, 705-708.

Dadlez, M. (1997). Hydrophobic interactions accelerate early stages of the folding of BPTI. *Biochemistry, 36*, 2788-2797.

Daggett, V. & Fersht, A. R. (2003). Is there a unifying mechanism for protein folding? *Trends Biochem Sci, 28*, 18-25.

de Groot, N. S., Aviles, F. X., 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.

de Groot, N. S., Sabate, R. & Ventura, S. (2009). Amyloids in bacterial inclusion bodies. *Trends Biochem Sci, 34*, 408-416.

de Groot, N. S. & Ventura, S. (2006). Protein activity in bacterial inclusion bodies correlates with predicted aggregation rates. *J Biotechnol, 125*, 110-113.

Dinner, A. R. & Karplus, M. (1998). A metastable state in folding simulations of a protein model. *Nat Struct Biol, 5*, 236-241.

Dobson, C. M. (2003). Protein folding and misfolding. *Nature, 426*, 884-890.

Dobson, C. M. (2004). Principles of protein folding, misfolding and aggregation. *Semin Cell Dev Biol, 15*, 3-16.

Doig, A. J. & Williams, D. H. (1992). Why water-soluble, compact, globular proteins have similar specific enthalpies of unfolding at 110 degrees C. *Biochemistry, 31*, 9371-9375.

Doye, J. P., Louis, A. A. & Vendruscolo, M. (2004). Inhibition of protein crystallization by evolutionary negative design. *Phys Biol, 1*, P9-13.

DuBay, K. F., Pawar, A. P., Chiti, F., Zurdo, J., Dobson, C. M. & Vendruscolo, M. (2004). Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains. *J Mol Biol, 341*, 1317-1326.

Dumoulin, M., Canet, D., Last, A. M., Pardon, E., Archer, D. B., Muyldermans, S., Wyns, L., Matagne, A., Robinson, C. V., Redfield, C. & Dobson, C. M. (2005). Reduced global cooperativity is a common feature underlying the amyloidogenicity of pathogenic lysozyme mutations. *J Mol Biol, 346*, 773-788.

Dumoulin, M., Last, A. M., Desmyter, A., Decanniere, K., Canet, D., Larsson, G., Spencer, A., Archer, D. B., Sasse, J., Muyldermans, S., Wyns, L., Redfield, C., Matagne, A., Robinson, C. V. & Dobson, C. M. (2003). A camelid antibody fragment inhibits the formation of amyloid fibrils by human lysozyme. *Nature, 424*, 783-788.

Eaglestone, S. S., Cox, B. S. & Tuite, M. F. (1999). Translation termination efficiency can be regulated in Saccharomyces cerevisiae by environmental stress through a prion-mediated mechanism. *EMBO J, 18*, 1974-1981.

Eichmann, C., Preissler, S., Riek, R. & Deuerling, E. (2010). Cotranslational structure acquisition of nascent polypeptides monitored by NMR spectroscopy. *Proc Natl Acad Sci U S A, 107*, 9111-9116.

Elam, J. S., Taylor, A. B., Strange, R., Antonyuk, S., Doucette, P. A., Rodriguez, J. A., Hasnain, S. S., Hayward, L. J., Valentine, J. S., Yeates, T. O. & Hart, P. J. (2003). Amyloid-like filaments and water-filled nanotubes formed by SOD1 mutant proteins linked to familial ALS. *Nat Struct Biol, 10*, 461-467.

Eulitz, M., Weiss, D. T. & Solomon, A. (1990). Immunoglobulin heavy-chain-associated amyloidosis. *Proc Natl Acad Sci U S A, 87*, 6542-6546.

Fernandez-Escamilla, A. M., 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.

Fernandez-Recio, J., Totrov, M. & Abagyan, R. (2004). Identification of protein-protein interaction sites from docking energy landscapes. *J Mol Biol, 335*, 843-865.

Fernandez-Recio, J., Totrov, M., Skorodumov, C. & Abagyan, R. (2005). Optimal docking area: a new method for predicting protein-protein interaction sites. *Proteins, 58*, 134-143.

Fersht, A. R. (1995). Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. *Proc Natl Acad Sci U S A, 92*, 10869-10873.

Fersht, A. R. (1997). Nucleation mechanisms in protein folding. *Curr Opin Struct Biol, 7*, 3-9.

Fink, A. L. (1998). Protein aggregation: folding aggregates, inclusion bodies and amyloid. *Fold Des, 3*, R9-23.

Folkers, P. J., Clore, G. M., Driscoll, P. C., Dodt, J., Kohler, S. & Gronenborn, A. M. (1989). Solution structure of recombinant hirudin and the Lys-47----Glu mutant: a nuclear magnetic resonance and hybrid distance geometry-dynamical simulated annealing study. *Biochemistry, 28*, 2601-2617.

Foss, T. R., Wiseman, R. L. & Kelly, J. W. (2005). The pathway by which the tetrameric protein transthyretin dissociates. *Biochemistry, 44*, 15525-15533.

Galzitskaya, O. V., Garbuzynskiy, S. O. & Lobanov, M. Y. (2006a). FoldUnfold: web server for the prediction of disordered regions in protein chain. *Bioinformatics, 22*, 2948-2949.

Galzitskaya, O. V., Garbuzynskiy, S. O. & Lobanov, M. Y. (2006b). Prediction of amyloidogenic and disordered regions in protein chains. *PLoS Comput Biol, 2*, e177.

Garcia-Fruitos, E., Carrio, M. M., Aris, A. & Villaverde, A. (2005a). Folding of a misfolding-prone beta-galactosidase in absence of DnaK. *Biotechnol Bioeng, 90*, 869-875.

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

Goldstrohm, A. C., Albrecht, T. R., Sune, C., Bedford, M. T. & Garcia-Blanco, M. A. (2001). The transcription elongation factor CA150 interacts with RNA polymerase II and the pre-mRNA splicing factor SF1. *Mol Cell Biol, 21*, 7617-7628.

Gonzalez-Montalban, N., Villaverde, A. & Aris, A. (2007). Amyloid-linked cellular toxicity triggered by bacterial inclusion bodies. *Biochem Biophys Res Commun, 355*, 637-642.

Gosal, W. S., Morten, I. J., Hewitt, E. W., Smith, D. A., Thomson, N. H. & Radford, S. E. (2005). Competing pathways determine fibril morphology in the self-assembly of beta2-microglobulin into amyloid. *J Mol Biol, 351*, 850-864.

Grana-Montes, R., de Groot, N. S., Castillo, V., Sancho, J., Velazquez-Campoy, A. & Ventura, S. Contribution of disulfide bonds to stability, folding, and amyloid fibril formation: the PI3-SH3 domain case. *Antioxid Redox Signal, 16*, 1-15.

Grantcharova, V. P., Riddle, D. S., Santiago, J. V. & Baker, D. (1998). Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. *Nat Struct Biol, 5*, 714-720.

Grutter, M. G., Priestle, J. P., Rahuel, J., Grossenbacher, H., Bode, W., Hofsteenge, J. & Stone, S. R. (1990). Crystal structure of the thrombin-hirudin complex: a novel mode of serine protease inhibition. *EMBO J, 9***,** 2361-2365.

Guerois, R., Nielsen, J. E. & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol, 320***,** 369-387.

Hartl, F. U. & Hayer-Hartl, M. (2002). Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science, 295***,** 1852-1858.

Hoyer, W., Gronwall, C., Jonsson, A., Stahl, S. & Hard, T. (2008). Stabilization of a beta-hairpin in monomeric Alzheimer's amyloid-beta peptide inhibits amyloid formation. *Proc Natl Acad Sci U S A, 105***,** 5099-5104.

Hurle, M. R., Helms, L. R., Li, L., Chan, W. & Wetzel, R. (1994). A role for destabilizing amino acid replacements in light-chain amyloidosis. *Proc Natl Acad Sci U S A, 91***,** 5446-5450.

Itzhaki, L. S., Neira, J. L., Ruiz-Sanz, J., de Prat Gay, G. & Fersht, A. R. (1995). Search for nucleation sites in smaller fragments of chymotrypsin inhibitor 2. *J Mol Biol, 254***,** 289-304.

Ivanova, M. I., Sawaya, M. R., Gingery, M., Attinger, A. & Eisenberg, D. (2004). An amyloid-forming segment of beta2-microglobulin suggests a molecular model for the fibril. *Proc Natl Acad Sci U S A, 101***,** 10584-10589.

Jahn, T. R., Parker, M. J., Homans, S. W. & Radford, S. E. (2006). Amyloid formation under physiological conditions proceeds via a native-like folding intermediate. *Nat Struct Mol Biol, 13***,** 195-201.

Jahn, T. R. & Radford, S. E. (2005). The Yin and Yang of protein folding. *FEBS J, 272***,** 5962-5970.

Janin, J., Miller, S. & Chothia, C. (1988). Surface, subunit interfaces and interior of oligomeric proteins. *J Mol Biol, 204***,** 155-164.

Jemth, P., Day, R., Gianni, S., Khan, F., Allen, M., Daggett, V. & Fersht, A. R. (2005). The structure of the major transition state for folding of an FF domain from experiment and simulation. *J Mol Biol, 350***,** 363-378.

Jemth, P., Gianni, S., Day, R., Li, B., Johnson, C. M., Daggett, V. & Fersht, A. R. (2004). Demonstration of a low-energy on-pathway intermediate in a fast-folding protein by kinetics, protein engineering, and simulation. *Proc Natl Acad Sci U S A, 101***,** 6450-6455.

Jemth, P., Johnson, C. M., Gianni, S. & Fersht, A. R. (2008). Demonstration by burst-phase analysis of a robust folding intermediate in the FF domain. *Protein Eng Des Sel, 21***,** 207-214.

Jones, S. & Thornton, J. M. (1996). Principles of protein-protein interactions. *Proc Natl Acad Sci U S A, 93***,** 13-20.

Jones, S. & Thornton, J. M. (1997a). Analysis of protein-protein interaction sites using surface patches. *J Mol Biol, 272***,** 121-132.

Jones, S. & Thornton, J. M. (1997b). Prediction of protein-protein interaction sites using patch analysis. *J Mol Biol, 272*, 133-143.

Kameda, A., Hoshino, M., Higurashi, T., Takahashi, S., Naiki, H. & Goto, Y. (2005). Nuclear magnetic resonance characterization of the refolding intermediate of beta2-microglobulin trapped by non-native prolyl peptide bond. *J Mol Biol, 348*, 383-397.

Karplus, M. & Weaver, D. L. (1976). Protein-folding dynamics. *Nature, 260*, 404-406.

Kayed, R., Sokolov, Y., Edmonds, B., McIntire, T. M., Milton, S. C., Hall, J. E. & Glabe, C. G. (2004). Permeabilization of lipid bilayers is a common conformation-dependent activity of soluble amyloid oligomers in protein misfolding diseases. *J Biol Chem, 279*, 46363-46366.

Kim, P. S. & Baldwin, R. L. (1982). Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. *Annu Rev Biochem, 51*, 459-489.

Kim, P. S. & Baldwin, R. L. (1990). Intermediates in the folding reactions of small proteins. *Annu Rev Biochem, 59*, 631-660.

Korzhnev, D. M., Religa, T. L., Lundstrom, P., Fersht, A. R. & Kay, L. E. (2007). The folding pathway of an FF domain: characterization of an on-pathway intermediate state under folding conditions by (15)N, (13)C(alpha) and (13)C-methyl relaxation dispersion and (1)H/(2)H-exchange NMR spectroscopy. *J Mol Biol, 372*, 497-512.

Korzhnev, D. M., Vernon, R. M., Religa, T. L., Hansen, A. L., Baker, D., Fersht, A. R. & Kay, L. E. Nonnative interactions in the FF domain folding pathway from an atomic resolution structure of a sparsely populated intermediate: an NMR relaxation dispersion study. *J Am Chem Soc, 133*, 10974-10982.

Kuhlman, B. & Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proc Natl Acad Sci U S A, 97*, 10383-10388.

Kumar, S. & Walter, J. (2011). Phosphorylation of amyloid beta (Abeta) peptides - a trigger for formation of toxic aggregates in Alzheimer's disease. *Aging (Albany NY), 3*, 803-812.

Lai, Z., Colon, W. & Kelly, J. W. (1996). The acid-mediated denaturation pathway of transthyretin yields a conformational intermediate that can self-assemble into amyloid. *Biochemistry, 35*, 6470-6482.

Lawrence, M. C. & Colman, P. M. (1993). Shape complementarity at protein/protein interfaces. *J Mol Biol, 234*, 946-950.

Lee, S., Sowa, M. E., Choi, J. M. & Tsai, F. T. (2004). The ClpB/Hsp104 molecular chaperone-a protein disaggregating machine. *J Struct Biol, 146*, 99-105.

Levinthal, C. (1968). Are there pathways for protein folding?. *J Med Phys, 65*, 44-45.

Liemann, S. & Glockshuber, R. (1999). Influence of amino acid substitutions related to inherited human prion diseases on the thermodynamic stability of the cellular prion protein. *Biochemistry, 38*, 3258-3267.

Lin, S. H., Konishi, Y., Denton, M. E. & Scheraga, H. A. (1984). Influence of an extrinsic cross-link on the folding pathway of ribonuclease A. Conformational and thermodynamic analysis of cross-linked (lysine7-lysine41)-ribonuclease a. *Biochemistry, 23*, 5504-5512.

Lindner, A. B., Madden, R., Demarez, A., Stewart, E. J. & Taddei, F. (2008). Asymmetric segregation of protein aggregates is associated with cellular aging and rejuvenation. *Proc Natl Acad Sci U S A, 105*, 3076-3081.

Lu, M., Yang, J., Ren, Z., Sabui, S., Espejo, A., Bedford, M. T., Jacobson, R. H., Jeruzalmi, D., McMurray, J. S. & Chen, X. (2009). Crystal structure of the three tandem FF domains of the transcription elongation regulator CA150. *J Mol Biol, 393*, 397-408.

Mamathambika, B. S. & Bardwell, J. C. (2008). Disulfide-linked protein folding pathways. *Annu Rev Cell Dev Biol, 24*, 211-235.

Martinez, J. C., Pisabarro, M. T. & Serrano, L. (1998). Obligatory steps in protein folding and the conformational diversity of the transition state. *Nat Struct Biol, 5*, 721-729.

Martinez, J. C. & Serrano, L. (1999). The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nat Struct Biol, 6*, 1010-1016.

Martinez, J. C., Viguera, A. R., Berisio, R., Wilmanns, M., Mateo, P. L., Filimonov, V. V. & Serrano, L. (1999). Thermodynamic analysis of alpha-spectrin SH3 and two of its circular permutants with different loop lengths: discerning the reasons for rapid folding in proteins. *Biochemistry, 38*, 549-559.

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

Mayer, S., Rudiger, S., Ang, H. C., Joerger, A. C. & Fersht, A. R. (2007). Correlation of levels of folded recombinant p53 in escherichia coli with thermodynamic stability in vitro. *J Mol Biol, 372*, 268-276.

Miller, S. (1989). The structure of interfaces between subunits of dimeric and tetrameric proteins. *Protein Eng, 3*, 77-83.

Monsellier, E. & Chiti, F. (2007). Prevention of amyloid-like aggregation as a driving force of protein evolution. *EMBO Rep, 8*, 737-742.

Monsellier, E., Ramazzotti, M., de Laureto, P. P., Tartaglia, G. G., Taddei, N., Fontana, A., Vendruscolo, M. & Chiti, F. (2007). The distribution of residues in a polypeptide sequence is a determinant of aggregation optimized by evolution. *Biophys J, 93*, 4382-4391.

Monsellier, E., Ramazzotti, M., Taddei, N. & Chiti, F. (2008). Aggregation propensity of the human proteome. *PLoS Comput Biol, 4*, e1000199.

Montelione, G. T., Wuthrich, K., Burgess, A. W., Nice, E. C., Wagner, G., Gibson, K. D. & Scheraga, H. A. (1992). Solution structure of murine epidermal growth factor

determined by NMR spectroscopy and refined by energy minimization with restraints. *Biochemistry, 31*, 236-249.

Morel, B., Varela, L., Azuaga, A. I. & Conejero-Lara, F. (2010). Environmental conditions affect the kinetics of nucleation of amyloid fibrils and determine their morphology. *Biophys J, 99*, 3801-3810.

Morell, M., Bravo, R., Espargaro, A., Sisquella, X., Aviles, F. X., Fernandez-Busquets, X. & Ventura, S. (2008). Inclusion bodies: specificity in their aggregation process and amyloid-like structure. *Biochim Biophys Acta, 1783*, 1815-1825.

Morris, D. P. & Greenleaf, A. L. (2000). The splicing factor, Prp40, binds the phosphorylated carboxyl-terminal domain of RNA polymerase II. *J Biol Chem, 275*, 39935-39943.

Murakami, Y. & Jones, S. (2006). SHARP2: protein-protein interaction predictions using patch analysis. *Bioinformatics, 22*, 1794-1795.

Musacchio, A., Noble, M., Pauptit, R., Wierenga, R. & Saraste, M. (1992). Crystal structure of a Src-homology 3 (SH3) domain. *Nature, 359*, 851-855.

Naiki, H., Gejyo, F. & Nakakuki, K. (1997). Concentration-dependent inhibitory effects of apolipoprotein E on Alzheimer's beta-amyloid fibril formation in vitro. *Biochemistry, 36*, 6243-6250.

Nakayashiki, T., Kurtzman, C. P., Edskes, H. K. & Wickner, R. B. (2005). Yeast prions [URE3] and [PSI+] are diseases. *Proc Natl Acad Sci U S A, 102*, 10575-10580.

Narayan, M., Welker, E., Wedemeyer, W. J. & Scheraga, H. A. (2000). Oxidative folding of proteins. *Acc Chem Res, 33*, 805-812.

Nelson, R., Sawaya, M. R., Balbirnie, M., Madsen, A. O., Riekel, C., Grothe, R. & Eisenberg, D. (2005). Structure of the cross-beta spine of amyloid-like fibrils. *Nature, 435*, 773-778.

Oberg, K., Chrunyk, B. A., Wetzel, R. & Fink, A. L. (1994). Nativelike secondary structure in interleukin-1 beta inclusion bodies by attenuated total reflectance FTIR. *Biochemistry, 33*, 2628-2634.

Ogiso, H., Ishitani, R., Nureki, O., Fukai, S., Yamanaka, M., Kim, J. H., Saito, K., Sakamoto, A., Inoue, M., Shirouzu, M. & Yokoyama, S. (2002). Crystal structure of the complex of human epidermal growth factor and receptor extracellular domains. *Cell, 110*, 775-787.

Otzen, D. E., Kristensen, O. & Oliveberg, M. (2000). Designed protein tetramer zipped together with a hydrophobic Alzheimer homology: a structural clue to amyloid assembly. *Proc Natl Acad Sci U S A, 97*, 9907-9912.

Pace, C. N., Grimsley, G. R., Thomson, J. A. & Barnett, B. J. (1988). Conformational stability and activity of ribonuclease T1 with zero, one, and two intact disulfide bonds. *J Biol Chem, 263*, 11820-11825.

Pantoja-Uceda, D., Arolas, J. L., Aviles, F. X., Santoro, J., Ventura, S. & Sommerhoff, C. P. (2009). Deciphering the structural basis that guides the oxidative folding of leech-derived tryptase inhibitor. *J Biol Chem, 284*, 35612-35620.

Pawar, A. P., Dubay, K. F., Zurdo, J., Chiti, F., Vendruscolo, M. & Dobson, C. M. (2005). Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases. *J Mol Biol, 350*, 379-392.

Pedersen, J. S., Christensen, G. & Otzen, D. E. (2004). Modulation of S6 fibrillation by unfolding rates and gatekeeper residues. *J Mol Biol, 341*, 575-588.

Periole, X., Vendruscolo, M. & Mark, A. E. (2007). Molecular dynamics simulations from putative transition states of alpha-spectrin SH3 domain. *Proteins, 69*, 536-550.

Petkova, A. T., Ishii, Y., Balbach, J. J., Antzutkin, O. N., Leapman, R. D., Delaglio, F. & Tycko, R. (2002). A structural model for Alzheimer's beta -amyloid fibrils based on experimental constraints from solid state NMR. *Proc Natl Acad Sci U S A, 99*, 16742-16747.

Plakoutsi, G., Bemporad, F., Calamai, M., Taddei, N., Dobson, C. M. & Chiti, F. (2005). Evidence for a mechanism of amyloid formation involving molecular reorganisation within native-like precursor aggregates. *J Mol Biol, 351*, 910-922.

Plakoutsi, G., Bemporad, F., Monti, M., Pagnozzi, D., Pucci, P. & Chiti, F. (2006). Exploring the mechanism of formation of native-like and precursor amyloid oligomers for the native acylphosphatase from Sulfolobus solfataricus. *Structure, 14*, 993-1001.

Plakoutsi, G., Taddei, N., Stefani, M. & Chiti, F. (2004). Aggregation of the Acylphosphatase from Sulfolobus solfataricus: the folded and partially unfolded states can both be precursors for amyloid formation. *J Biol Chem, 279*, 14111-14119.

Plaza del Pino, I. M., Ibarra-Molero, B. & Sanchez-Ruiz, J. M. (2000). Lower kinetic limit to protein thermal stability: a proposal regarding protein stability in vivo and its relation with misfolding diseases. *Proteins, 40*, 58-70.

Poland, D.C., Scheraga, H.A. (1965). Statistical mechanics of noncovalent bonds in polyamino acids. VIII. Covalent loops in proteins. *Biopolymers, 3*, 379–399.

Quintas, A., Saraiva, M. J. & Brito, R. M. (1999). The tetrameric protein transthyretin dissociates to a non-native monomer in solution. A novel model for amyloidogenesis. *J Biol Chem, 274*, 32943-32949.

Quintas, A., Vaz, D. C., Cardoso, I., Saraiva, M. J. & Brito, R. M. (2001). Tetramer dissociation and monomer partial unfolding precedes protofibril formation in amyloidogenic transthyretin variants. *J Biol Chem, 276*, 27207-27213.

Ray, S. S. & Lansbury, P. T., Jr. (2004). A possible therapeutic target for Lou Gehrig's disease. *Proc Natl Acad Sci U S A, 101*, 5701-5702.

Ren, L., McLean, J. R., Hazbun, T. R., Fields, S., Vander Kooi, C., Ohi, M. D. & Gould, K. L. Systematic two-hybrid and comparative proteomic analyses reveal novel yeast pre-mRNA splicing factors connected to Prp19. *PLoS One, 6*, e16719.

Reumers, J., Maurer-Stroh, S., Schymkowitz, J. & Rousseau, F. (2009). Protein sequences encode safeguards against aggregation. *Hum Mutat, 30*, 431-437.

Richardson, J. S. & Richardson, D. C. (2002). Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc Natl Acad Sci U S A, 99*, 2754-2759.

Ritter, C., Maddelein, M. L., Siemer, A. B., Luhrs, T., Ernst, M., Meier, B. H., Saupe, S. J. & Riek, R. (2005). Correlation of structural elements and infectivity of the HET-s prion. *Nature, 435*, 844-848.

Rousseau, F., Serrano, L. & Schymkowitz, J. W. (2006). How evolutionary pressure against protein aggregation shaped chaperone specificity. *J Mol Biol, 355*, 1037-1047.

Routledge, K. E., Tartaglia, G. G., Platt, G. W., Vendruscolo, M. & Radford, S. E. (2009). Competition between intramolecular and intermolecular interactions in an amyloid-forming protein. *J Mol Biol, 389*, 776-786.

Rydel, T. J., Ravichandran, K. G., Tulinsky, A., Bode, W., Huber, R., Roitsch, C. & Fenton, J. W., 2nd (1990). The structure of a complex of recombinant hirudin and human alpha-thrombin. *Science, 249*, 277-280.

Sabate, R., Espargaro, A., de Groot, N. S., Valle-Delgado, J. J., Fernandez-Busquets, X. & Ventura, S. (2010). The role of protein sequence and amino acid composition in amyloid formation: scrambling and backward reading of IAPP amyloid fibrils. *J Mol Biol, 404*, 337-352.

Sadqi, M., Casares, S., Abril, M. A., Lopez-Mayorga, O., Conejero-Lara, F. & Freire, E. (1999). The native state conformational ensemble of the SH3 domain from alpha-spectrin. *Biochemistry, 38*, 8899-8906.

Salamanca, S., Li, L., Vendrell, J., Aviles, F. X. & Chang, J. Y. (2003). Major kinetic traps for the oxidative folding of leech carboxypeptidase inhibitor. *Biochemistry, 42*, 6754-6761.

Salamanca, S., Villegas, V., Vendrell, J., Li, L., Aviles, F. X. & Chang, J. Y. (2002). The unfolding pathway of leech carboxypeptidase inhibitor. *J Biol Chem, 277*, 17538-17543.

Sanchez-Ruiz, J. M. (1992). Theoretical analysis of Lumry-Eyring models in differential scanning calorimetry. *Biophys J, 61*, 921-935.

Sanchez de Groot, N., Pallares, I., Aviles, F. X., Vendrell, J. & Ventura, S. (2005). Prediction of "hot spots" of aggregation in disease-linked polypeptides. *BMC Struct Biol, 5*, 18.

Sanchorawala, V. (2006). Light-chain (AL) amyloidosis: diagnosis and treatment. *Clin J Am Soc Nephrol, 1*, 1331-1341.

Saupe, S. J. (2000). Molecular genetics of heterokaryon incompatibility in filamentous ascomycetes. *Microbiol Mol Biol Rev, 64*, 489-502.

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

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F. & Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Res, 33*, W382-388.

Sekijima, Y., Wiseman, R. L., Matteson, J., Hammarstrom, P., Miller, S. R., Sawkar, A. R., Balch, W. E. & Kelly, J. W. (2005). The biological and chemical basis for tissue-selective amyloid disease. *Cell, 121*, 73-85.

Serag, A. A., Altenbach, C., Gingery, M., Hubbell, W. L. & Yeates, T. O. (2002). Arrangement of subunits and ordering of beta-strands in an amyloid sheet. *Nat Struct Biol, 9*, 734-739.

Serio, T. R., Cashikar, A. G., Kowal, A. S., Sawicki, G. J., Moslehi, J. J., Serpell, L., Arnsdorf, M. F. & Lindquist, S. L. (2000). Nucleated conformational conversion and the replication of conformational information by a prion determinant. *Science, 289*, 1317-1321.

Soldi, G., Bemporad, F., Torrassa, S., Relini, A., Ramazzotti, M., Taddei, N. & Chiti, F. (2005). Amyloid formation of a protein in the absence of initial unfolding and destabilization of the native state. *Biophys J, 89*, 4234-4244.

Tartaglia, G. G., Cavalli, A., Pellarin, R. & Caflisch, A. (2004). The role of aromaticity, exposed surface, and dipole moment in determining protein aggregation rates. *Protein Sci, 13*, 1939-1941.

Tartaglia, G. G., Cavalli, A. & Vendruscolo, M. (2007). Prediction of local structural stabilities of proteins from their amino acid sequences. *Structure, 15*, 139-143.

Tartaglia, G. G. & Vendruscolo, M. (2008). The Zyggregator method for predicting protein aggregation propensities. *Chem Soc Rev, 37*, 1395-1401.

Thompson, M. J., Sievers, S. A., Karanicolas, J., Ivanova, M. I., Baker, D. & Eisenberg, D. (2006). The 3D profile method for identifying fibril-forming segments of proteins. *Proc Natl Acad Sci U S A, 103*, 4074-4078.

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.

True, H. L. & Lindquist, S. L. (2000). A yeast prion provides a mechanism for genetic variation and phenotypic diversity. *Nature, 407*, 477-483.

Uversky, V. N., E, M. C., Bower, K. S., Li, J. & Fink, A. L. (2002). Accelerated alpha-synuclein fibrillation in crowded milieu. *FEBS Lett, 515*, 99-103.

Vakser, I. A. & Aflalo, C. (1994). Hydrophobic docking: a proposed enhancement to molecular recognition techniques. *Proteins, 20*, 320-329.

Ventura, S., Vega, M. C., Lacroix, E., Angrand, I., Spagnolo, L. & Serrano, L. (2002). Conformational strain in the hydrophobic core and its implications for protein folding and design. *Nat Struct Biol, 9*, 485-493.

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

Viguera, A. R., Jimenez, M. A., Rico, M. & Serrano, L. (1996). Conformational analysis of peptides corresponding to beta-hairpins and a beta-sheet that represent the entire sequence of the alpha-spectrin SH3 domain. *J Mol Biol, 255*, 507-521.

Viguera, A. R., Martinez, J. C., Filimonov, V. V., Mateo, P. L. & Serrano, L. (1994). Thermodynamic and kinetic analysis of the SH3 domain of spectrin shows a two-state folding transition. *Biochemistry, 33*, 2142-2150.

Wang, M.C., Uhlenbeck, G.E. (1945). On the theory of the Brownian motion II. *Rev Mod Phys, 17*, 323–342.

Wang, L., Maji, S. K., Sawaya, M. R., Eisenberg, D. & Riek, R. (2008). Bacterial inclusion bodies contain amyloid-like structure. *PLoS Biol, 6*, e195.

Wedemeyer, W. J., Welker, E., Narayan, M. & Scheraga, H. A. (2000). Disulfide bonds and protein folding. *Biochemistry, 39*, 4207-4216.

Weissman, J. S. & Kim, P. S. (1991). Reexamination of the folding of BPTI: predominance of native intermediates. *Science, 253*, 1386-1393.

Weissman, J. S. & Kim, P. S. (1992). Kinetic role of nonnative species in the folding of bovine pancreatic trypsin inhibitor. *Proc Natl Acad Sci U S A, 89*, 9900-9904.

Westermark, P., Benson, M. D., Buxbaum, J. N., Cohen, A. S., Frangione, B., Ikeda, S., Masters, C. L., Merlini, G., Saraiva, M. J. & Sipe, J. D. (2007). A primer of amyloid nomenclature. *Amyloid, 14*, 179-183.

Wetlaufer, D. B. (1973). Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc Natl Acad Sci U S A, 70*, 697-701.

Wickner, S., Maurizi, M. R. & Gottesman, S. (1999). Posttranslational quality control: folding, refolding, and degrading proteins. *Science, 286*, 1888-1893.

Yoon, S. & Welsh, W. J. (2004). Detecting hidden sequence propensity for amyloid fibril formation. *Protein Sci, 13*, 2149-2160.

Young, L., Jernigan, R. L. & Covell, D. G. (1994). A role for surface hydrophobicity in protein-protein recognition. *Protein Sci, 3*, 717-729.

Zavodszky, M., Chen, C. W., Huang, J. K., Zolkiewski, M., Wen, L. & Krishnamoorthi, R. (2001). Disulfide bond effects on protein stability: designed variants of Cucurbita maxima trypsin inhibitor-V. *Protein Sci, 10*, 149-160.

Zibaee, S., Makin, O. S., Goedert, M. & Serpell, L. C. (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.

Zwanzig, R., Szabo, A. & Bagchi, B. (1992). Levinthal's paradox. *Proc Natl Acad Sci U S A, 89*, 20-22.

VII - SCIENTIFIC PAPERS

doi:10.1016/j.jmb.2009.06.049 *J. Mol. Biol.* (2009) **392**, 529–546

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Designing Out Disulfide Bonds of Leech Carboxypeptidase Inhibitor: Implications for Its Folding, Stability and Function

Joan L. Arolas, Virginia Castillo, Sílvia Bronsoms, Francesc X. Aviles* and Salvador Ventura*

Institut de Biotecnologia i Biomedicina, Universitat Autònoma de Barcelona, E-08193 Bellaterra, Barcelona, Spain

Departament de Bioquímica i Biologia Molecular, Facultat de Ciències, Universitat Autònoma de Barcelona, E-08193 Bellaterra, Barcelona, Spain

Received 13 January 2009; received in revised form 4 May 2009; accepted 18 June 2009
Available online 25 June 2009

Leech carboxypeptidase inhibitor (LCI) is a 67-residue, tight-binding metallocarboxypeptidase inhibitor composed of a compact domain with a five-stranded β-sheet and a short α-helix that are strongly stabilized by four disulfide bonds. In this study, we investigated the contribution of each particular disulfide to the folding, stability and function of LCI by constructing a series of single and multiple mutants lacking one to four disulfide bonds. The results allow a better understanding of how individual disulfide bonds shape and restrict the conformational space that LCI must explore before attaining its native conformation. The work also dissected the role played by intramolecular rearrangements of disulfides during LCI folding, providing a new kinetic scheme in which the 2S ensemble suffers a non-specific oxidation into the 3S ensemble. These 3-disulfide-bonded species reshuffle to preferentially form III-A and III-B, two major native-like folding intermediates that need structural rearrangements through the formation of scrambled isomers to finally render native LCI. The designed multiple mutants of LCI are unable to fold correctly, displaying a highly unstructured conformation and a very low inhibitory capability, which indicates the importance of disulfide bonds in LCI for both correct folding and achievement of a functional structure. In contrast, the elimination of a single disulfide bond in LCI only results in a significant reduction of conformational stability, but the mutations have a rather moderate impact on carboxypeptidase inhibition, allowing the possibility to target the intrinsic stability and specific activity of LCI independently. In this way, the findings reported provide a basis for the design of novel variants of the molecule with improved therapeutic properties.

Keywords: metallocarboxypeptidase inhibitor; disulfide bond; protein folding; protein stability; protein engineering

Edited by F. Schmid

**Corresponding authors.* F. X. Aviles is to be contacted at Protein Engineering Unit, Universitat Autònoma de Barcelona, E-08193 Bellaterra, Barcelona, Spain. S. Ventura, Protein Folding and Design Unit, Universitat Autònoma de Barcelona, E-08193 Bellaterra, Barcelona, Spain. E-mail addresses: francescxavier.aviles@uab.es; salvador.ventura@uab.es.

Abbreviations used: bis-ANS, 4,4′-bis(1-anilinonaphthalene-8-sulfonate); BPTI, bovine pancreatic trypsin inhibitor; CPA, carboxypeptidase A; GdnHCl, guanidine hydrochloride; GSH, reduced glutathione; GSSG, oxidized glutathione; K_i, inhibition constant; LCI, leech carboxypeptidase inhibitor; MALDI-TOF MS, matrix-assisted laser desorption/ionization time-of-flight mass spectrometry; MCP, metallocarboxypeptidase; RP-HPLC, reversed-phase high-performance liquid chromatography; TFA, trifluoroacetic acid; WT, wild type.

Introduction

The covalent link of cysteine residues by disulfide bonds is vital for the folding, stability and function of many proteins both in bacteria and in eukaryotes and is a topic of wide interest in the fields of protein folding and protein engineering.[1–3] The removal of a disulfide bond by reduction or substitution for another amino acid residue usually results in a significant conformational destabilization of the protein.[4–8] In addition, it may have a positive or negative effect on the folding rate and folding efficiency of the protein by narrowing the folding landscape or by leading to kinetically trapped intermediates during the folding reaction.[9–12] The term "oxidative folding" describes the composite process by which a reduced unfolded protein gains both its native disulfide bonds (disulfide-bond formation) and its native structure (conformational folding).[13,14] Numerous studies directed to investigate the oxi-

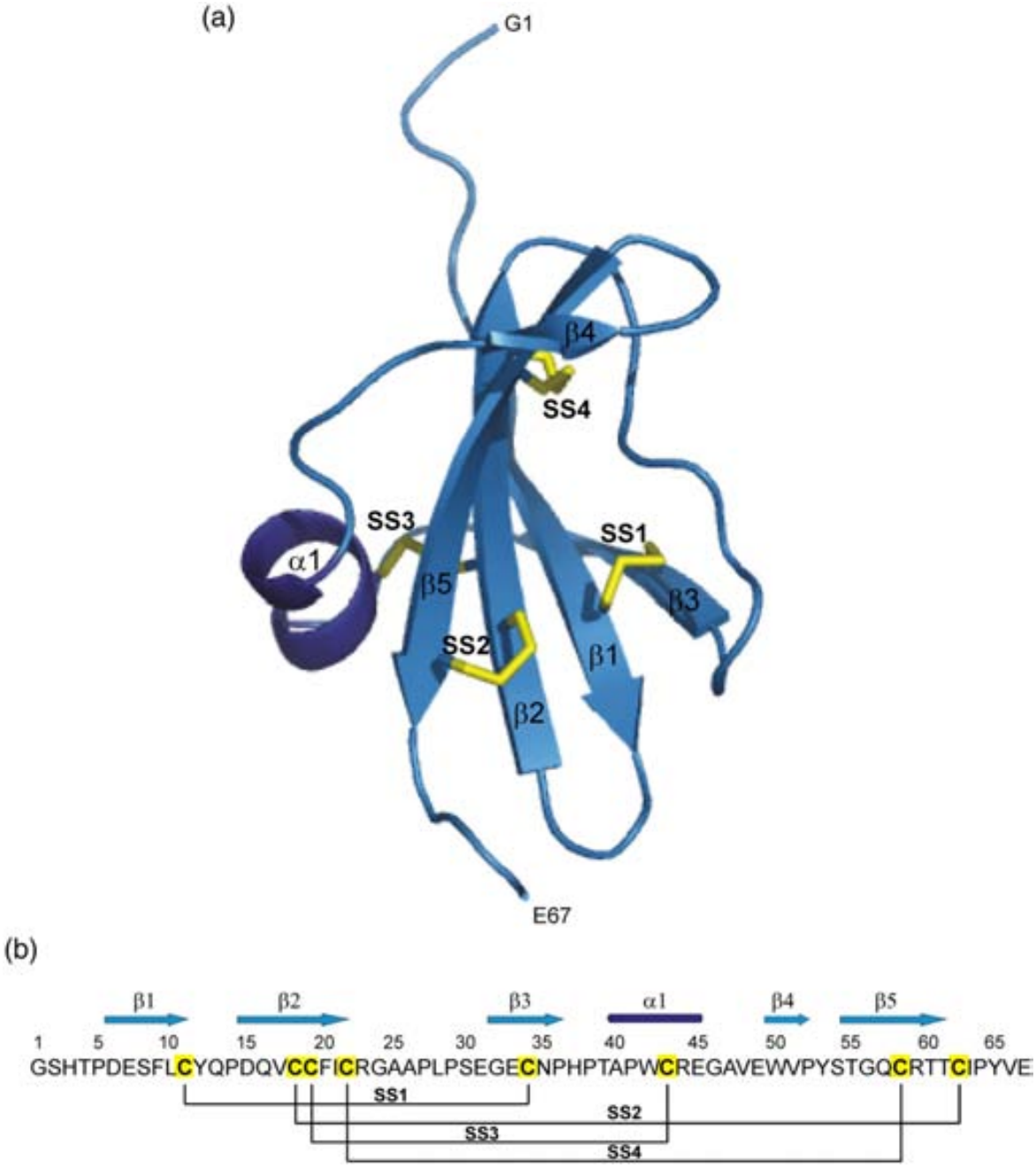


Fig. 1. Structure of LCI. (a) Ribbon representation of the three-dimensional structure of native LCI. The four native disulfide bonds (SS1, Cys11–Cys34; SS2, Cys18–Cys62; SS3, Cys19–Cys43; and SS4, Cys22–Cys58) are shown in yellow sticks. The N- and C-terminal residues are labeled. The Protein Data Bank accession code for the structure is 1DTV. This figure was prepared with PyMOL. (b) Amino acid sequence of LCI. Its secondary structure elements and disulfide-bond pairings are schematically shown above and below the sequence, respectively.

dative folding of small disulfide-rich proteins have taken advantage of the particular chemistry of disulfide-bond formation.[15–17] In those studies, the proteins are initially fully reduced/unfolded and then allowed to refold in the presence of selected buffers containing different redox agents. The folding intermediates that arise during the folding reaction are subsequently trapped in a time-course manner by acidification (reversible) or alkylation (irreversible) and separated by reversed-phase high-performance liquid chromatography (RP-HPLC), which makes their further disulfide-bond and structural characterization possible. The extent of heterogeneity of folding intermediates, the predominance of intermediates containing native disulfide bonds and native (sub)structures and the presence of scrambled isomers (fully oxidized species that contain at least two non-native disulfide bonds) as intermediates are used to define the mechanism of oxidative folding of the analyzed protein.[18] Surprisingly, a great diversity of folding mechanisms have been observed among small disulfide-rich proteins, even for those containing similar disulfide-bond patterns and three-dimensional structures.[19,20] At the two extremes of this diversity we find bovine pancreatic trypsin inhibitor (BPTI) and hirudin, two well-known serine protease inhibitors. BPTI folds through a limited number of disulfide intermediates adopting native disulfide pairings and native-like structures that funnel the folding reaction toward the native protein and prevent the accumulation of scrambled isomers.[21,22] In contrast, the folding of hirudin shows a high degree of heterogeneity of intermediates, including an extensive population of scrambled isomers that act as major kinetic traps slowing the folding reaction.[23–25] Other proteins, such as leech carboxypeptidase inhibitor (LCI), display both similarities and dissimilarities to the folding of BPTI and hirudin, owing to the presence of both native-like intermediates and scrambled isomers.[26,27]

LCI is a 67-residue cysteine-rich protein isolated from the medical leech *Hirudo medicinalis* that inhibits metallocarboxypeptidases (MCPs) of the A/B subfamily (M14A, according to the MEROPS database) with nanomolar affinity.[28] The three-dimensional structure of LCI shows that it folds in a compact domain consisting of a five-stranded antiparallel β-sheet and a short α-helix that are strongly stabilized by the presence of four disulfide bonds (Fig. 1).[29] LCI is a potent inhibitor of plasma carboxypeptidase B, also known as CPU or thrombin-activatable fibrinolysis inhibitor, a molecular link between blood coagulation and fibrinolysis.[30,31] Assuming that leeches secrete LCI during feeding, the inhibitor could help maintain the liquid state of blood by inhibition of thrombin-activatable fibrinolysis inhibitor. Indeed, our group has recently tested the *in vitro* profibrinolytic activity of LCI and demonstrated its possible use as an enhancer of the tissue-type plasminogen activator therapy in thrombotic disorders (S. Salamanca *et al.*, unpublished data). Knowledge about the folding and unfolding determinants of this molecule is essential for the development of variants with enhanced stability and/or altered activity. We have previously described the unfolding pathway and conformational stability of LCI, revealing that this protein has slow unfolding kinetics and is highly stable against denaturation.[32] We have also extensively characterized its oxidative folding pathway, showing a rapid and sequential flow of 1- and 2-disulfide intermediates that reaches a rate-limiting step in which two 3-disulfide intermediates (termed III-A and III-B) act as major kinetic traps.[33] These two intermediates contain only native disulfide bonds (III-A and III-B lack Cys22–Cys58 and Cys19–Cys43, respectively) and need major structural rearrangements through the formation of a heterogeneous population of 4-disulfide scrambled isomers to render native LCI. We have recently determined the nuclear magnetic resonance (NMR) structure of trapped and isolated III-A,[34] as well as the crystal structure of an analog of III-B,[35] reporting that both intermediates are metastable and contain a native-like structure that is responsible for their strong accumulation during the LCI folding reaction. They display greater flexibility as compared with the native form, especially at the regions surrounding the free cysteines (or alanines in the analog), however. In this study, we examined the effect of disulfide-bond removal on the folding, stability and function of LCI using site-directed mutagenesis to replace each cystine with a pair of alanine residues.

Results

To study the contribution of the four disulfide bonds to the folding, stability and function of LCI, we constructed seven mutant proteins in which single or multiple disulfide bonds were disrupted by Cys→Ala mutations (Table 1). They include the mutants lacking one of the four native disulfide bonds of LCI (LCI_1 to LCI_4), a double mutant lacking the two native disulfide bonds absent in the III-A and III-B intermediates of wild-type (WT) LCI folding (LCI_3/4), a triple mutant displaying only the Cys11–Cys34 disulfide bond (LCI_2/3/4) and a mutant lacking the four native disulfide bonds of the inhibitor (LCI_1/2/3/4). The mutant and WT forms were expressed using *Escherichia coli* cells harboring a plasmid containing the OmpA signal peptide for

Table 1. LCI mutant proteins constructed in this work

Name	Missing disulfide bond(s)	Mutations
LCI_1	SS1	C11A/C34A
LCI_2	SS2	C18A/C62A
LCI_3	SS3	C19A/C43A
LCI_4	SS4	C22A/C58A
LCI_3/4	SS3, SS4	C19A/C43A, C22A/C58A
LCI_2/3/4	SS2, SS3, SS4	C18A/C62A, C19A/C43A, C22A/C58A
LCI_1/2/3/4	SS1, SS2, SS3, SS4	C11A/C34A, C18A/C62A, C19A/C43A, C22A/C58A

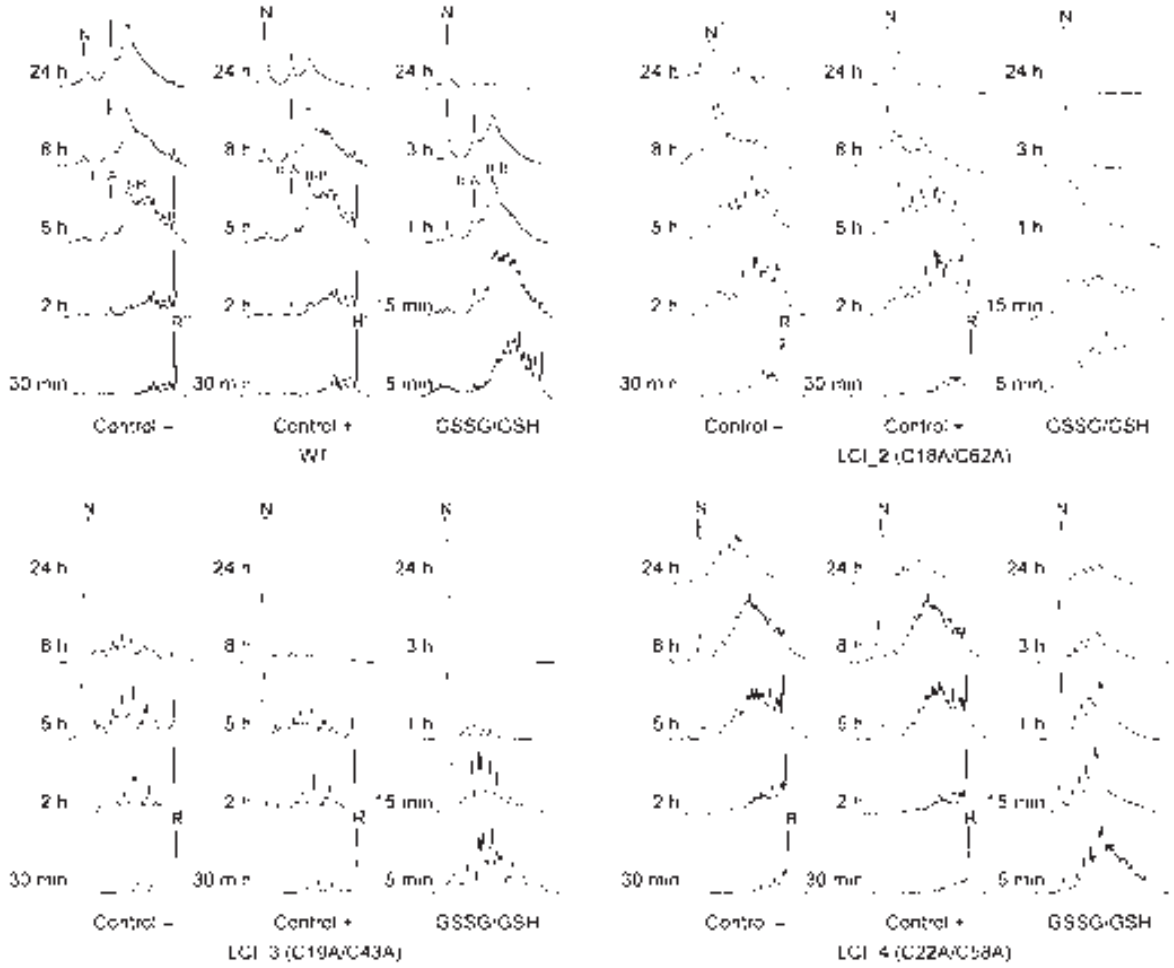


Fig. 2. Oxidative folding of WT and LCI_2, LCI_3 and LCI_4 mutants. RP-HPLC traces of the acid-trapped intermediates that occur along the oxidative folding of different LCI variants. The reactions were carried out in Tris–HCl buffer, pH 8.5, in the absence (Control −) and in the presence (Control +) of 0.25 mM 2-mercaptoethanol or a mixture of GSSG and GSH (0.5 and 1 mM) as detailed in Materials and Methods. The retention times of the native (N) and fully reduced/unfolded (R) forms are labeled. III-A and III-B are two native 3-disulfide intermediates of the oxidative folding of WT LCI.

periplasmic expression and were purified by a combination of three chromatography steps that rendered soluble proteins with purity greater than 95% as described in Materials and Methods. About 3 mg/l of protein was obtained for WT LCI. However, the expression yields of the single and multiple LCI mutants were much lower (10–50 times) depending on the identity of the mutated disulfide bond: WT>LCI_3>LCI_2>LCI_4>LCI_1>LCI_3/4>LCI_2/3/4~LCI_1/2/3/4.

Oxidative folding of WT and LCI mutants

Fully reduced and unfolded WT, LCI_1, LCI_2, LCI_3, LCI_4 and LCI_3/4 forms were allowed to refold in Tris–HCl buffer, pH 8.5, in the absence (Control −) and in the presence (Control +) of either 2-mercaptoethanol or a mixture of oxidized glutathione and reduced glutathione (GSSG and GSH). The intermediates arising along the folding reaction of the WT and mutant proteins were trapped by acidification or alkylation at selected time points and further analyzed by RP-HPLC and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) to investigate their chromatographic behavior and disulfide-bond content, respectively. A high degree of heterogeneity of intermediates was observed at the beginning of the folding process of WT LCI (up to 8 h), with similar RP-HPLC profiles observed

Fig. 3. Disulfide species in the oxidative folding of WT and LCI_2, LCI_3 and LCI_4 mutants. Folding was performed in Tris–HCl buffer, pH 8.5, in the absence (Control −) and in the presence (Control +) of 0.25 mM 2-mercaptoethanol or a mixture of GSSG and GSH (0.5 and 1 mM). The percentages of disulfide species were determined by alkylation with 4-vinylpyridine and subsequent analysis by MALDI-TOF MS as detailed in Materials and Methods. *XS* represents an ensemble of species with an *X* number of disulfide bonds. The recovery of the native LCI variants is represented by bars and was calculated from the peak areas in the corresponding RP-HPLC chromatograms.

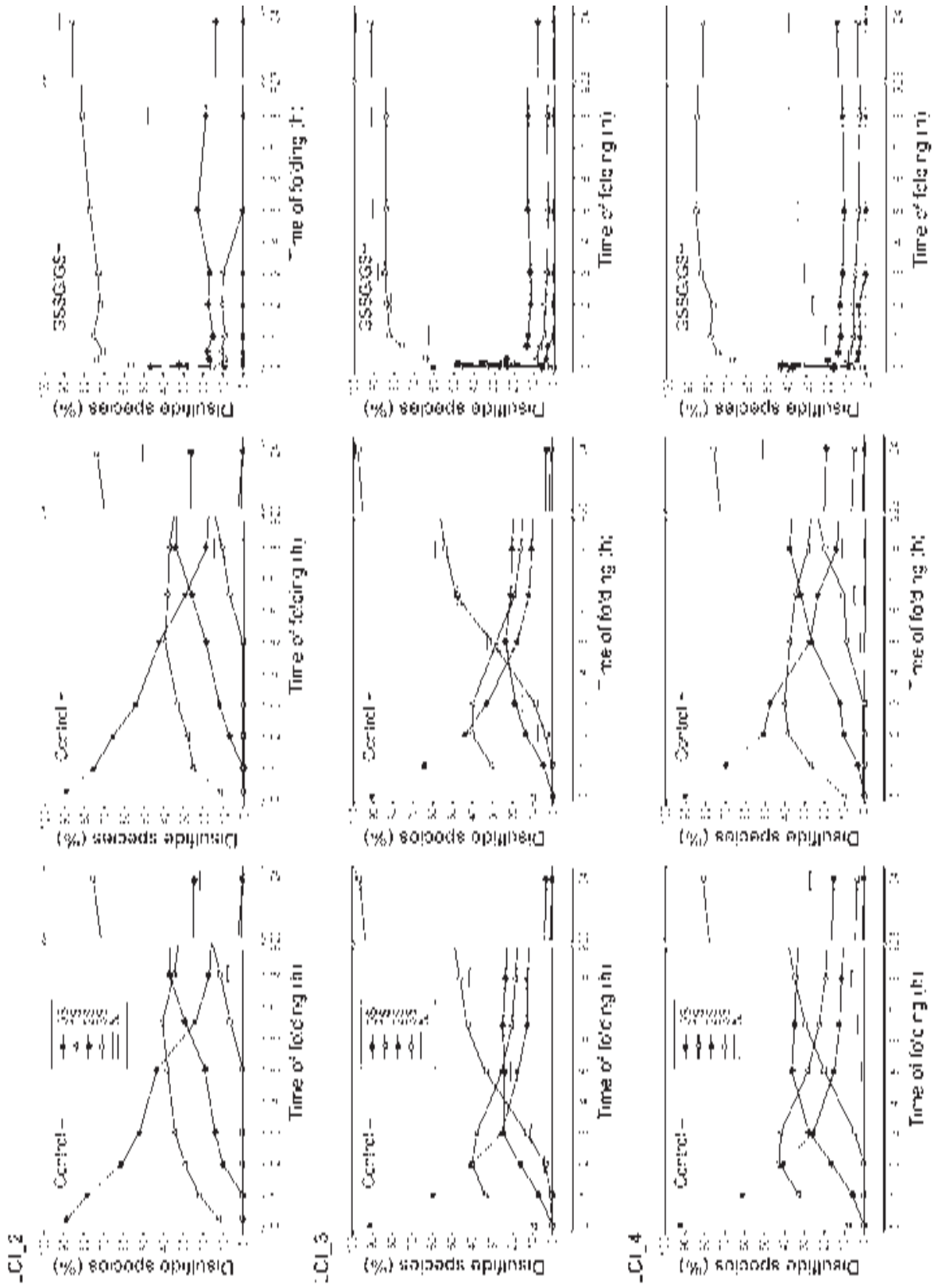


Fig. 3 *(legend on previous page)*

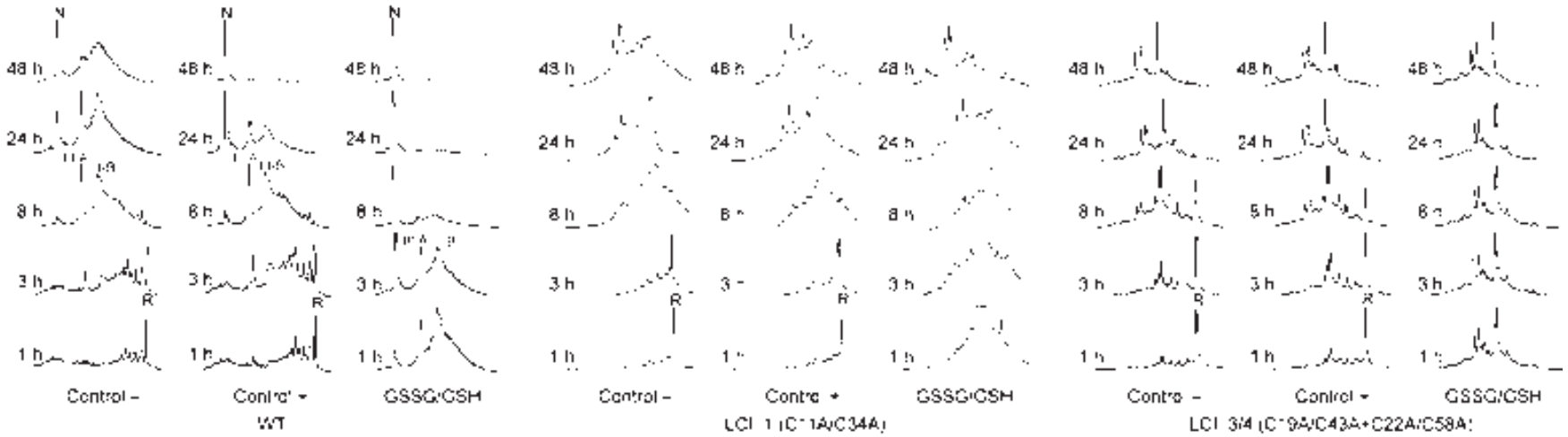


Fig. 4. Oxidative folding of LCI_1 and LCI_3/4 mutants. RP-HPLC traces of the acid-trapped intermediates that occur along the oxidative folding of different LCI variants. The reactions were carried out in Tris–HCl buffer, pH 8.5, in the absence (Control −) and in the presence (Control +) of 0.25 mM 2-mercaptoethanol or a mixture of GSSG and GSH (0.5 and 1 mM). The retention times of the native (N) and fully reduced/unfolded (R) forms are labeled. III-A and III-B are two native 3-disulfide intermediates of the oxidative folding of WT LCI.

regardless of the presence of thiol catalyst (see Fig. 2, Control − and Control +). The folding reaction was shown to undergo a sequential flow through the formation of 1-, 2- and 3-disulfide intermediates that ended up with a strong accumulation of two native 3-disulfide intermediates, III-A and III-B (Fig. 2). The last step of folding was characterized by the occurrence of a heterogeneous population of 4-disulfide intermediates that needed the presence of 2-mercaptoethanol or GSH to reshuffle their non-native disulfide bonds and efficiently render native protein. The addition of oxidizing reagent (GSSG) strongly accelerated the folding rate, promoting the coexistence of both 3-disulfide intermediates (III-A and III-B) and a mixture of 4-disulfide scrambled isomers as major kinetic traps of the reaction.

Of the five LCI mutants that were analyzed in terms of oxidative folding, only LCI_2, LCI_3 and LCI_4 were able to acquire a native-like state (Fig. 2), although they followed different folding mechanisms. LCI_2 and LCI_4 folded through heterogeneous populations of 1- and 2-disulfide intermediates, leading to the accumulation of a complex population of 3-disulfide scrambled isomers that slowed the folding rate (Fig. 2). As in the case of WT LCI, only the presence of thiol catalyst could promote disulfide-bond rearrangements toward the native disulfide-bond pairing and native-like structure. However, for LCI_4, the folding process was not very efficient and only approximately 50% of the protein was recovered as native form after treatment with GSSG/GSH for 24 h (Fig. 3). In contrast, LCI_3 rapidly folded through smaller populations of 1- and 2-disulfide intermediates and a low percentage of 3-disulfide scrambled isomers, as indicated by the slight effect exhibited by the addition of thiol catalyst (Figs. 2 and 3). Again, the presence of oxidizing reagent (GSSG) strongly speeded the folding rate by acceleration of disulfide-bond formation. The absence of a strong kinetic barrier in the folding reaction of this mutant (i.e., the reshuffling of scrambled isomers into the native protein) accounted for an extremely efficient process with more than 90% of native protein recovered after 24 h in the absence of redox agents (Control −).

LCI_1 and LCI_3/4 could not reach any predominant conformation, not even after incubation with GSSG/GSH for 48 h (Fig. 4). LCI_1 folding underwent a sequential formation of 1-, 2- and 3-disulfide intermediates, leading to the accumulation of a highly heterogeneous population of scrambled isomers that could not rearrange into any native-like state. The case of the double mutant was similar to that of LCI_1 and only differed in the amount of occurring intermediates, which were more reduced owing to the lower number of possible disulfide species. The initial formation of 1-disulfide intermediates was followed by the final accumulation of three major 2-disulfide scrambled isomers that could not evolve into any predominant form (see GSSG/GSH condition in Fig. 4). The expression of LCI_2/3/4 and LCI_1/2/3/4, containing one and no disulfide bond, respectively, rendered a unique

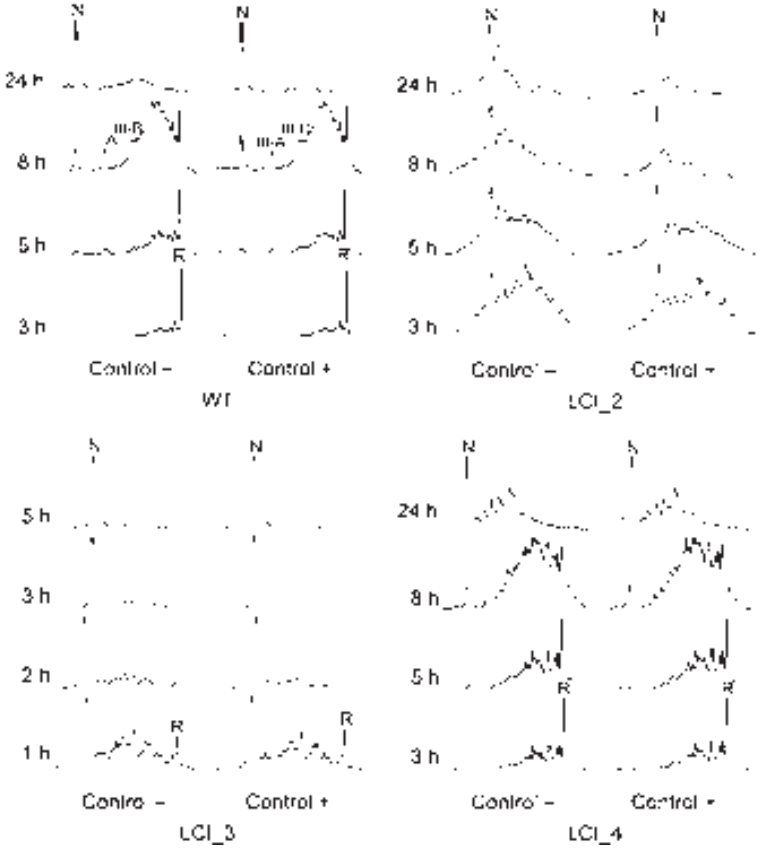


Fig. 5. Effect of denaturant on the oxidative folding of WT and LCI_2, LCI_3 and LCI_4 mutants. RP-HPLC traces of the acid-trapped intermediates that occur along the oxidative folding of different LCI variants. The reactions were carried out in Tris–HCl buffer, pH 8.5, containing 0.5 M GdnHCl in the absence (Control −) and in the presence (Control +) of 0.25 mM 2-mercaptoethanol. The retention times of the native (N) and fully reduced/unfolded (R) forms are labeled. III-A and III-B are two native 3-disulfide intermediates of the oxidative folding of WT LCI.

conformation. Nevertheless, both mutants eluted in the RP-HPLC in a position close to that of the fully reduced/unfolded form of WT LCI (data not shown), suggesting a high degree of hydrophobicity and low compactness. On the other hand, the native-like forms of LCI_2, LCI_3 and LCI_4 eluted at time points located between those of the native and fully reduced/unfolded forms of WT LCI, pointing to a medium degree of hydrophobicity and compactness.

As previously reported,[33] the addition of low concentrations of denaturant strongly accelerates the folding of WT LCI by diminishing the accumulation of the native-like intermediates III-A and III-B, which act as kinetic traps of the reaction. The oxidative folding of the three LCI mutants that could achieve a native-like conformation, LCI_2, LCI_3 and LCI_4, was examined using 0.5 M guanidine hydrochloride (GdnHCl) both in the absence (Control −) and in the presence (Control +) of thiol catalyst. While the folding of LCI_2 and LCI_4 was slightly affected by the addition of the denaturant (Fig. 5), the formation of native LCI_3 was strongly accelerated by its presence.

Reductive unfolding and disulfide scrambling of WT and LCI mutants

The reductive unfolding of the native form of WT, LCI_2, LCI_3 and LCI_4 was studied in Tris–HCl buffer, pH 8.5, using increasing concentrations of dithiothreitol (DTT) as reducing agent to test the stability of their disulfide bonds.[36,37] The mixtures were trapped in a time-course manner by acidification and further analyzed by RP-HPLC. Reduction of WT LCI underwent an apparent all-or-none mechanism in which only low amounts of III-A and III-B intermediates were detected along the reaction (Fig. 6). These two species were subsequently reduced into fully reduced LCI without a significant accumulation of 2- and 1-disulfide intermediates. The same reduction behavior was observed at different concentrations of DTT (1–100 mM), with variations only in the rates of the reduction process. The analysis also revealed an all-or-none reduction mechanism for the 3-disulfide LCI mutants, which however resisted much lower concentrations of DTT than the WT. The dependence of the fraction of reduced protein on the reaction time could be fitted to a single constant first-order reaction ($R>0.99$) for LCI_3 and LCI_4 (Fig. 7a and b). A linear dependence of the apparent reductive unfolding rates on the DTT concentration was observed (Fig. 7c), which allowed us to calculate average rate constants of 0.129 ± 0.002 and 0.406 ± 0.002 s^{-1} M^{-1}, for LCI_3 and LCI_4, respectively. The reduction rate of LCI_2 appeared to be in between those of LCI_3 and LCI_4 at any given DTT concentration, but the reactions could not be fitted to a first-order equation using a single constant, suggesting a different reduction mechanism for this mutant.

It is well established that in the presence of denaturant and thiol initiator, unfolding of a small disulfide-rich protein is accompanied by reshuffling of its native disulfide bonds, a process called disulfide scrambling.[38,39] Unfolding of native WT, LCI_2, LCI_3 and LCI_4 was carried out in Tris–HCl buffer, pH 8.5, in the presence of increasing concentrations of denaturant (urea or GdnHCl) and 0.25 mM 2-mercaptoethanol as thiol initiator. After reaching equilibrium for 20 h, the reactions were trapped by acidification and analyzed by RP-HPLC. The denaturation curves showed that the extent of unfolding resulting from the equilibrium between scrambled isomers and the native protein was clearly dependent upon the strength of denaturant used (Fig. 8a). Urea required in general higher concentrations than GdnHCl to denature the native proteins. The midpoint denaturant concentrations (c_M) were as follows: >8, 2 and 4.2 M for WT, LCI_2 and LCI_3, respectively, and 4.1, 1.0, 1.9 and 0.6 M GdnHCl for WT, LCI_2, LCI_3 and LCI_4, respectively. Thus, all the mutants are significantly less stable than the WT form (see thermodynamic parameters in Table 2). The relative stability of the

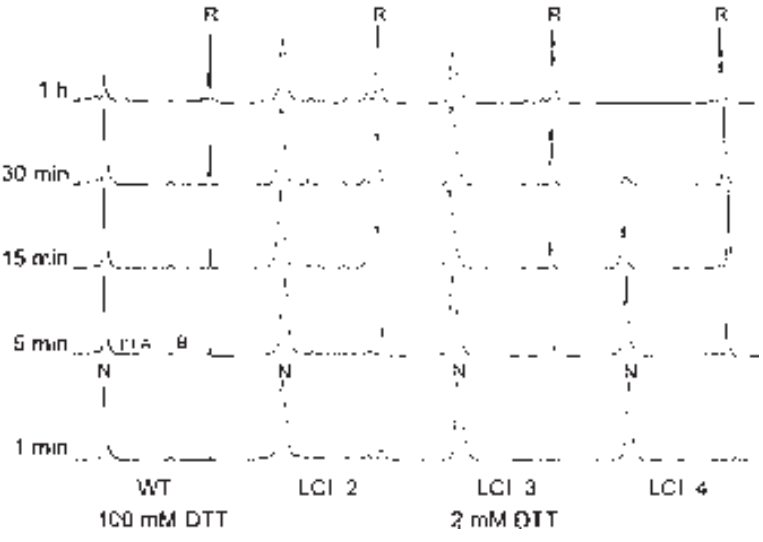


Fig. 6. Reductive unfolding of WT and LCI_2, LCI_3 and LCI_4 mutants. RP-HPLC traces of the acid-trapped intermediates that occur along the reductive unfolding of different LCI variants. The native proteins were treated with selected concentrations of DTT in Tris–HCl buffer, pH 8.5. "N" and "R" stand for the native and fully reduced/unfolded forms, respectively. III-A and III-B are two native 3-disulfide intermediates that also accumulate along the oxidative folding of WT LCI.

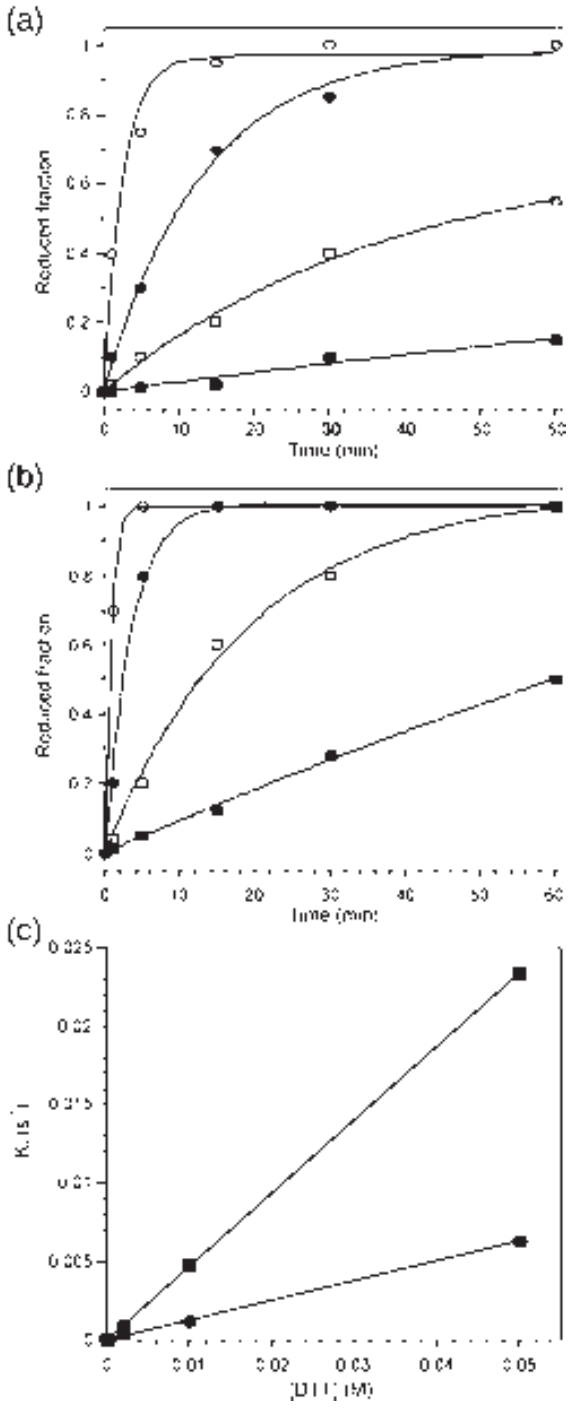


Fig. 7. Kinetics of reduction of LCI_3 and LCI_4 mutants. Reduction reactions for (a) LCI_3 and (b) LCI_4 were carried out in Tris–HCl buffer, pH 8.5, containing 0.5 mM (■), 2 mM (□), 10 mM (●) or 50 mM (○) DTT. The reactions were fitted to first-order reactions as described in Materials and Methods. (c) Linear dependence of the calculated reduction rates from DTT concentration for LCI_3 (●) and LCI_4 (■) mutants.

different LCI mutants is in agreement with their level of expression and with the percentages of scrambled isomers obtained after reaching equilibrium in the presence of thiol initiator and in the absence of denaturant (Fig. 8b). While the WT form and LCI_3 do not reshuffle into any scrambled isomer population, displaying structural uniqueness, LCI_2 and LCI_4 reshuffle into 10% and 45% of scrambled isomers, respectively, with K'_{eq} values of 0.1 and 0.45.

Conformational properties of WT and LCI mutants

The structural features of the native forms of WT and LCI mutants were analyzed by circular dichroism (CD) and NMR spectroscopy. The CD spectrum of the native WT LCI displays a well-defined minimum in ellipticity at 208 nm and a maximum at 228 nm, whose signal has been shown to correlate with the degree of unfolding of the protein.[35] The shape of the CD spectrum of LCI_3 is similar to that of WT LCI (Fig. 9a), except for a slight decrease in the intensity of the 228-nm maximum. The decrease in the intensity of this signature is more pronounced for LCI_4 and LCI_2, being especially significant the complete lack of the maximum and minimum of ellipticity of LCI_2/3/4 and LCI_1/2/3/4, which show a CD spectrum close to that of random-coil polypeptides. These results are further supported by NMR spectroscopy. The one-dimensional NMR spectrum of WT LCI displays a wide signal dispersion of resonances at both low (amide and aromatic region) and high (methyl region) fields, with good peak sharpness, characteristic of a folded molecule. The spectra of LCI_2, LCI_3 and LCI_4 show a lower dispersion of resonances at the amide region (Fig. 9b), mainly in the case of LCI_4, and those of LCI_2/3/4 and LCI_1/2/3/4 exhibit clear band broadening and peak collapse around 8 ppm. Altogether, these data indicate a partially folded conformation for the single mutants and a mostly unstructured conformation for the multiple mutants.

The conformation of LCI_1/2/3/4 was also assessed by fluorescence spectroscopy. The fluorescence spectrum of LCI is dominated by the contribution of two Tyr and two Trp residues (Tyr9, Tyr29, Trp42 and Trp50) buried in the interior of the protein that give an emission maximum at 350 nm (Fig. 9c).[35] The comparison of the spectrum of WT LCI with that of the multiple mutant reveals an unfolded conformation for the latter, which is supported by a red shift to 355–360 nm and a strong increase of intensity in the fluorescence emission. To test whether LCI_1/2/3/4 displays any residual structure, we incubated it with bis-ANS [4,4′-bis(1-anilinonaphthalene-8-sulfonate)]. The dye binds to hydrophobic surfaces from partially folded intermediates with much higher affinity than to native or completely unfolded proteins,[40] resulting in an increase and a blue shift of bis-ANS fluorescence emission. A 6-fold increase in bis-ANS fluorescence intensity and a strong shift toward lower wavelengths of the emission maximum were detected for LCI_1/2/3/4 (Fig. 9d), indicating the presence of hydrophobic clusters exposed to solvent. These

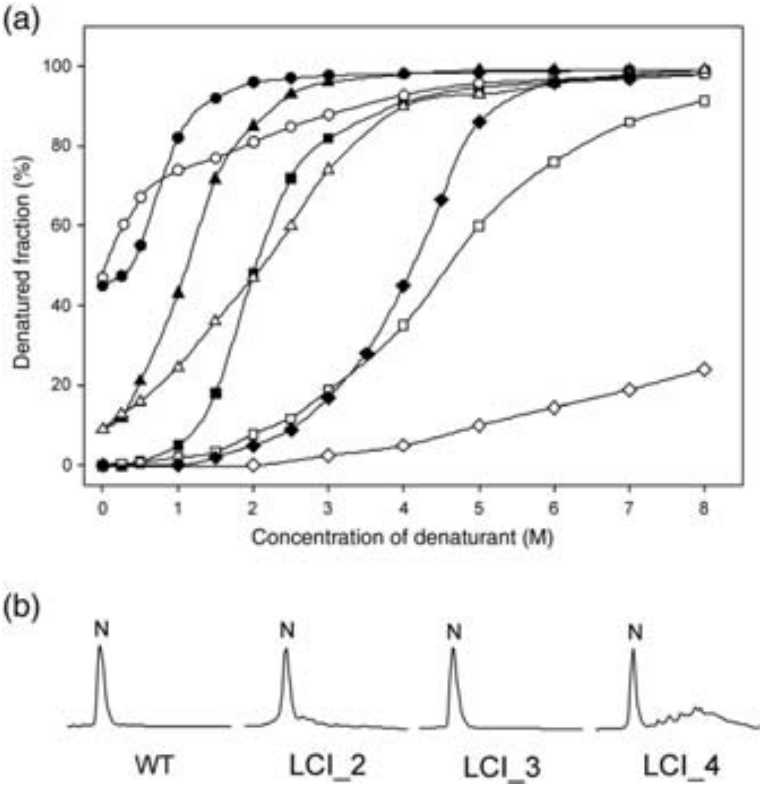


Fig. 8. Disulfide scrambling of WT and LCI_2, LCI_3 and LCI_4 mutants. (a) The denatured fraction of the different LCI variants was calculated as the percentage of native protein converted into scrambled isomers. WT LCI (◆), LCI_2 (▲), LCI_3 (■) and LCI_4 (●) were denatured in the presence of increasing concentrations of GdnHCl (filled symbols) or urea (open symbols) for 20 h at room temperature in Tris–HCl buffer, pH 8.5, containing 0.25 mM 2-mercaptoethanol as thiol initiator. (b) RP-HPLC traces of the acid-trapped scrambled isomers generated by disulfide scrambling. The native form of WT and LCI_2, LCI_3 and LCI_4 mutants was incubated for 20 h in Tris–HCl buffer, pH 8.5, containing only 0.25 mM 2-mercaptoethanol. The retention time of the native (N) forms is labeled.

hydrophobic surfaces are more protected in the WT structure as evidenced by its lower affinity for bis-ANS. Also, the bis-ANS signal of LCI_1/2/3/4 at increasing temperature exhibits a linear decrease without any detectable cooperativity (data not shown).

Inhibitory activity of WT and LCI mutants

WT LCI is a tight-binding competitive inhibitor of MCPs of the A/B subfamily with an inhibition constant (K_i) against bovine CPA (carboxypeptidase A; the archetypal MCP) in the nanomolar range (Table 3). The K_i values of the LCI mutants showed that LCI_2 and LCI_3 have an inhibitory potential similar to that of WT LCI. In contrast, LCI_4 suffered a significant loss of inhibitory activity with a K_i about 10 times higher than the WT value. LCI_2/3/4 and LCI_1/2/3/4 inhibited the bovine CPA with very low affinity, with a K_i in the micromolar range. On the other hand, long-term inhibitory assays conducted with LCI_3 revealed that this mutant displays a dissociation constant about 3 times higher than that in the WT form.

Table 2. Thermodynamic properties of unfolding of the WT and LCI single mutants

	ΔG (kcal mol^{-1})	m (kcal mol^{-1} M^{-1})	c_M (M)	$\Delta\Delta G_{WT-mut}$ (kcal mol^{-1})
LCI_2 urea	1.5±0.42	0.76±0.06	2.0	
LCI_3 urea	2.9±0.33	0.68±0.08	4.2	
WT GdnHCl	4.7±0.41	1.16±0.11	4.1	
LCI_2 GdnHCl	1.6±0.29	1.37±0.29	1.0	−3.1
LCI_3 GdnHCl	2.9±0.28	1.53±0.12	1.9	−1.8
LCI_4 GdnHCl	1.0±0.39	1.71±0.17	0.6	−3.7

The errors shown correspond to the fitting errors.

Discussion

Contribution of disulfide bonds to the conformational stability of LCI

One of the most intriguing properties of LCI folding is the role played by intramolecular rearrangements of disulfides during folding and how this correlates with the formation/disruption of its structure, conformational stability and function. LCI consists of a five-stranded antiparallel β-sheet with a β3–β1–β2–β5–β4 topology and a short α-helix that packs onto the most compact part of the β-structure. Three disulfide bonds (Cys11–Cys34, Cys18–Cys62 and Cys22–Cys58) sustain the structure of the β-sheet that is covalently connected to the α-helix by the Cys19–Cys43 disulfide bond. In this work, we have analyzed the stability of LCI mutants using the disulfide scrambling technique. The quantitative conversion of the native proteins to scrambled forms reflects their conformational stability and approaches to physiological conditions in which thiol catalysts are likely to be present. Elimination of disulfide bonds in LCI invariably results in a strong reduction in the conformational stability of the protein. Disulfide bonds are proposed to stabilize proteins by decreasing the conformational entropy of the denatured state.[41] The expected stabilization of the unfolded state and therefore the

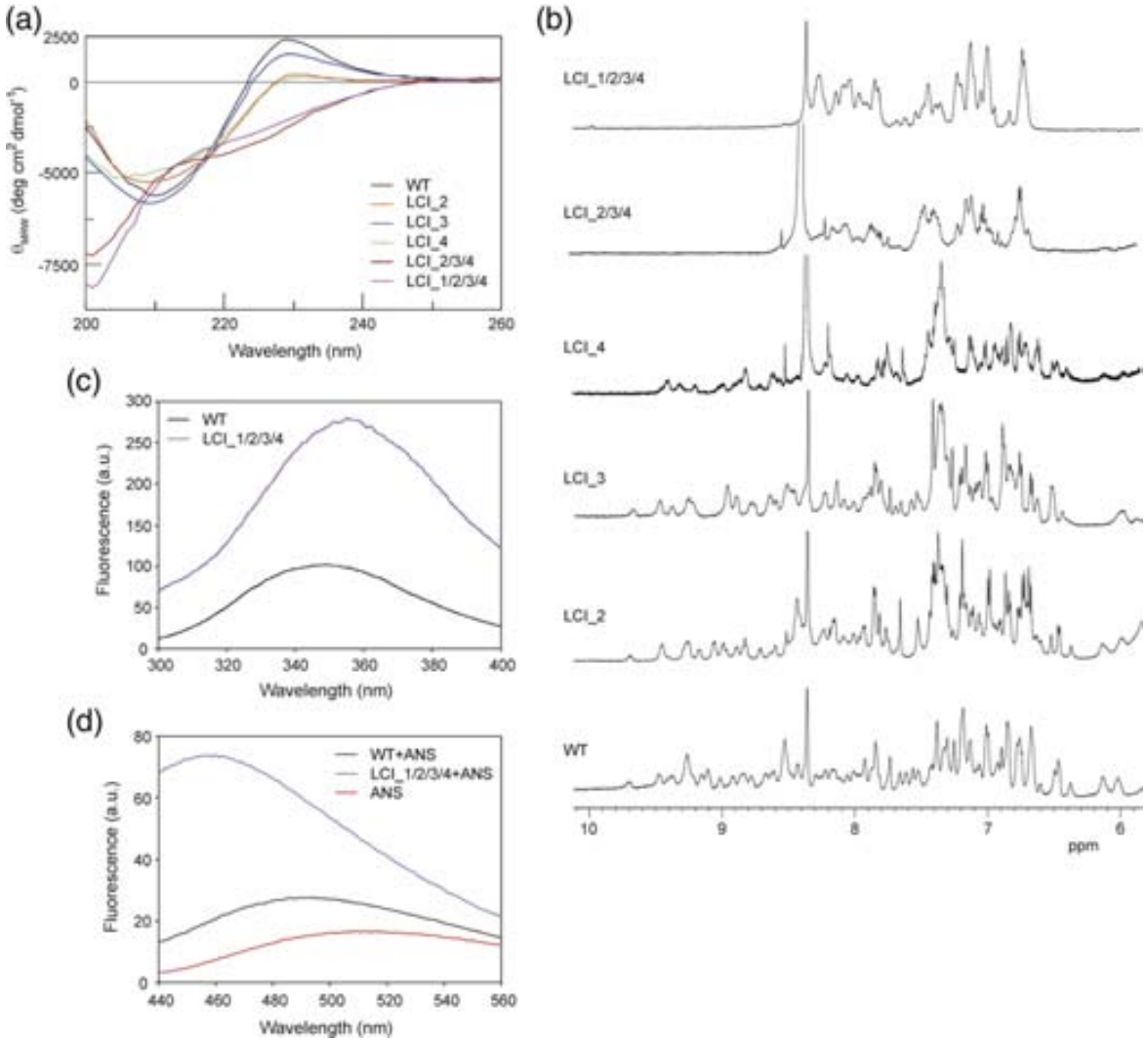


Fig. 9. Conformational properties of WT and LCI mutants. (a) Far-UV CD analyses of WT and LCI mutants were carried out at 25 °C in Tris–HCl buffer, pH 7.5, using a final protein concentration of 0.2 mg/ml. (b) One-dimensional NMR spectra of WT and LCI mutants were recorded at 25 °C and 600 MHz in Tris–HCl buffer, pH 7.5, using a protein concentration of 1.5 mg/ml. (c) Fluorescence emission analysis of WT and LCI_1/2/3/4 carried out at 25 °C in Tris–HCl buffer, pH 7.5, using 13 μM protein concentration. The excitation wavelength was set at 280 nm. (d) Fluorescence emission spectra of bis-ANS collected in the absence and in the presence of WT and LCI_1/2/3/4 at 25 °C. The excitation wavelength was set at 370 nm.

destabilization of a disulfide-containing protein upon removal of one or more covalent links can be calculated using the Wang–Uhlenbeck formula for loop entropy. [42] The predicted $\Delta\Delta G_{WT-mut}$ provides a framework to rationalize the effects of disulfide-bond removal on LCI conformation and folding.

Table 3. K_i values of the native forms of WT and LCI mutants against bovine CPA

	K_i (nM)
WT	1.1±0.2
LCI_2	2.1±0.3
LCI_3	2.4±0.3
LCI_4	12.6±0.5
LCI_2/3/4	3400±225
LCI_1/2/3/4	6500±240

The stability of WT LCI in the presence of a thiol agent is 4.7 kcal/mol. Mutants destabilized by a significant gain of loop entropy higher than this value are not expected to fold into a compact and unique structure. Accordingly, CD and NMR data clearly indicate that removal of more than two disulfide bonds in LCI results in unfolding toward mostly unstructured conformations in which the specific interactions among aromatic residues, characteristic of the native state of the inhibitor, are completely lost. In contrast to WT LCI, LCI_1/2/3/4 is able to significantly bind bis-ANS, indicating the presence of hydrophobic clusters exposed to solvent in this mutant. Nevertheless, the contacts sustaining these clusters appear to be non-specific since they are rapidly lost in a non-cooperative manner upon heating; also, the intrinsic fluorescence spectrum of Tyr and Trp residues indicates that their side chains

are fully accessible to solvent. Individual disulfide bonds SS3 (Cys19–Cys43) and SS4 (Cys22–Cys58) are missing in the major LCI folding intermediates (III-B and III-A, respectively), although they attain a folded conformation.[33] Loop entropy considerations predict that LCI_3/4 would be destabilized by about 6 kcal/mol and therefore would be unstructured, in spite of conserving the two disulfide bonds shared by the two major LCI folding intermediates. In this regard, the double mutant cannot fold into a preferential conformation and a heterogeneous population of 2S species is formed. The distribution of the LCI_3/4 2S species at equilibrium is independent of the presence of redox agents, which, as discussed in the next section, indicates that there is no preferential non-covalent interaction driving the formation of these species.

The predicted entropic destabilizations for the single LCI mutants suggest that they should be at least partially folded. In agreement with this prediction, we have shown that III-A and III-B intermediates display a native-like conformation and function. No stable intermediate lacking disulfide SS1 (Cys11–Cys34) or SS2 (Cys18–Cys62) has been detected to accumulate during the LCI folding reaction, and therefore their conformational properties are unknown. If only unfolded-state entropic considerations are taken into account, about 0.6 kcal/mol should separate the most and least stable mutants. However, the experimental data are in contradiction with this proposal: LCI_1 is unable to fold into a unique conformation and LCI_4 and LCI_3 are, respectively, less and more stable than they should be according to the entropic gain in their respective unfolded states. This indicates that apart from entropic factors, other forces influence the stability of single LCI mutants. To address this discrepancy, we have used the FoldX† force field developed by Guerois *et al.*[43] and Schymkowitz *et al.*[44] to predict stability changes in the single LCI mutants and the precise energy terms accounting for them. In this algorithm, the stabilization of the unfolded state by disulfide-bond removal is calculated using the logN rule of Darby and Creighton.[45] There is a good correlation between these values and those calculated with the Wang–Uhlenbeck approach (Table 4).

FoldX accurately predicts the rank and relative differences in stability for those LCI mutants able to fold into a predominant conformation (LCI_2, LCI_3 and LCI_4). A minor destabilization, relative to the WT form, is predicted for LCI_3 (−0.10 kcal/mol), whereas a striking loss of stability is predicted for LCI_4 (−4.17 kcal/mol). LCI_1 and LCI_2 are predicted to have intermediate stabilities (losses of −2.03 and −2.48 kcal/mol, respectively). Apart from the gain in entropy in the unfolded state, the main destabilizing effect appears to be a reduction of the van der Waals contributions, linked to the formation of cavities inside the protein due to the smaller volume of Ala side chains (Fig. 10). These destabilizing effects are partially compensated by the increase in entropy of the Ala side chains in the native state, relative to bonded Cys, in all the mutants. In LCI_3, according to the energy terms reported by the algorithm, the main stabilizing factor is a reduction of the energy penalization due to inter- and intraresidue steric overlaps relative to the WT protein, which suggests that the covalent link between the β-sheet and the α-helix introduces some strain in WT LCI. Importantly, the crystal structure of LCI_3 shows a relaxation of the backbone in the residues adjacent to the Ala mutations, especially in the last turn of the α-helix, which can significantly contribute to alleviate this strain.[35] The same effect, although to a minor extent, is predicted for the disulfide bonds connecting the highly regular lower part of the β-sheet. No such strain appears to exist in the less regular upper part of the β-sheet, and therefore the entropic destabilizing effect in LCI_4 cannot be effectively compensated. Accordingly, in the solution structure of the III-A mutant, lacking the Cys22–Cys58 disulfide bond, β-strands 2 and 5 connected by this bond adopt essentially the same structure as in the WT protein.[34]

The reduction rate for disulfide bonds buried in a stable folded conformation is nearly always found to be slow and directly involved in the rate-determining step of the reductive unfolding reaction.[37,46] With the exception of Cys62, all other Cys residues are deeply buried in the core of WT LCI, accounting for its high resistance to reduction. The corresponding cross-bridges would be shielded from reducing agents, such as DTT, except when conformational fluctuations would expose the disulfides. Thus, the lower conformational stability of single LCI mutants makes their disulfides much more reactive to DTT than the same bonds in the WT form. The linear dependence of reductive unfolding rates of LCI_3 and LCI_4 on the DTT concentration indicates that the rate-limiting step of their unfolding reaction is the reduction of the first disulfide bond. Accordingly, reduction of these mutants undergoes an apparent all-or-none mechanism in which all three disulfides are reduced in an apparent cooperative manner, with no significant accumulation of partially reduced intermediates. In excellent agreement with their relative conformational stabilities, LCI_4 is reduced 3.1 times faster than LCI_3, which confirms that this rate-limiting step is sensitive to

Table 4. Stabilization (ΔΔ*G*) of the unfolded state expected for each LCI mutant with respect to the WT at 25 °C based on loop entropy (in kcal mol^{-1})

	Wang–Uhlenbeck model	Creighton model
LCI_1	2.79	3.34
LCI_2	3.36	3.98
LCI_3	2.83	3.17
LCI_4	3.19	3.80
LCI_3/4	6.01	6.98
LCI_2/3/4	9.37	10.97
LCI_1/2/3/4	12.16	14.34

† http://foldx.crg.es/

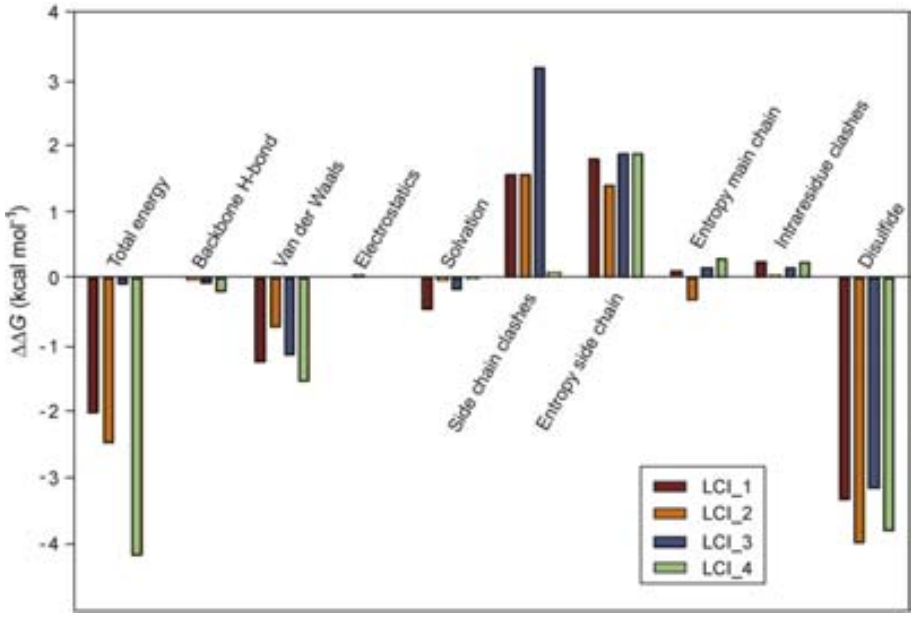


Fig. 10. FoldX predicted changes in stability for LCI_1, LCI_2, LCI_3 and LCI_4 mutants relative to WT LCI. The predicted overall change in stability ("Total energy") is the result of the contribution of the following: hydrogen-bond formation between backbone atoms ("Backbone H-bond"), van der Waals interactions ("van der Waals"), electrostatic interactions ("Electrostatics"), penalization for burying polar groups and solvation of hydrophobic groups ("Solvation"), clashes between atoms from different residues ("Side-chain clashes"), torsional intraresidue clashes ("Intraresidue clashes"), entropy cost of fixing side chains ("Entropy side chain"), entropy cost of fixing the backbone ("Entropy main chain") and entropy restriction in the unfolded state resulting from disulfide-bond formation/disruption ("Disulfide"). Other energy terms calculated by the algorithm but not affected by the mutations are not represented.

conformational changes in the protein. Interestingly, the reduction reaction of LCI_2, which lacks the most exposed Cys18–Cys62 disulfide bond, could not be fitted to a simple first-order reaction and the accumulation of a higher extent of partially reduced intermediates was detected, even at high DTT concentrations. This suggests that once reduced, the free thiolates of Cys18 and/or Cys62 might play a catalytic role in the concerted LCI disulfide reduction reaction.

An important result derived from the present study is that the relative stabilities of LCI_3 and LCI_4 mutants do not agree with those previously deduced for their III-B and III-A counterparts since III-A has been shown to be more stable and structured than III-B.[33] This reversion of the conformational properties of the major LCI folding intermediates responds to the effect of substituting free Cys in the intermediates by Ala in the correspondent analogs. Thus, the greater stability of III-A relative to LCI_4 may be due to the absence of a cavity in the hydrophobic core, provided that the thiols remain protonated. In agreement with this hypothesis, computation of Cys22 and Cys58 solvent accessibility with the DSSP‡ program indicates that they are totally buried in the solution structure of the intermediate. Conversely, the lower stability of III-B relative to LCI_3 may be due to charged thiolates preventing the packing of the α-helix onto the β-sheet. Unfortunately, no structure of the III-B intermediate is available. Mutations of these residues to Ser in the LCI_3 structure were modeled with FoldX and changes in stability were calculated to emulate the effect of polar residues in positions 19 and 43 of III-B. A 1.9-kcal/mol destabilization is predicted, which is in good agreement with the energetic cost of burying hydroxyl groups in the core of barnase[47] and suggests that the presence of charged thiolates would be strongly destabilizing. Cys-to-Ala mutations are considered the most neutral changes to emulate disulfide disruption in folding intermediates, especially when the disulfide bond is buried in the protein core. However, as deduced from our data, it is advisable to be cautious when attributing the conformational and thermodynamic properties of mutants to those of the correspondent intermediate forms without any other considerations, since they can be either overstabilized or destabilized relative to the des species depending on the accessibility, local environment or dihedral angles of the considered disulfide bond.

Role of disulfide bonds in the folding pathway of LCI

The data reported in the present and previous work of our group converge to indicate that the folding pathway of LCI hinges critically on the presence of localized stable structures. The different structural contents of III-B and III-A or LCI_3 and

‡ http://swift.cmbi.kun.nl/gv/dssp/

LCI_4 indicate that the accumulation of the major intermediates along the LCI folding reaction or the folding of their analogs into stable conformations relies on their ability to protect their three native disulfide bonds from rearrangement in the interior of totally or partially folded structures. In this context, the properties of LCI_2 constitute a paradox because, despite being conformationally stable, its des[18–62] counterpart does not accumulate during the oxidative folding reaction of WT LCI. This suggests that, as proposed for reductive unfolding, during refolding, the free thiolates Cys18 and Cys62 might attack other disulfide bonds quickly, causing the des[18–62] species to rearrange quickly to other disulfide species. This seems plausible since Cys62, the most exposed Cys residue in LCI, is found at the C-terminus and in an edge β-strand, two factors known to correlate with structural mobility. LCI_1 is predicted to have a structural stability similar to that of LCI_2. However, elimination of the Cys11–Cys34 disulfide bond in LCI_1 has the strongest impact on the LCI folding reaction. LCI_1 folding cannot be funneled into a single species with native disulfide connectivity, and the reaction becomes trapped in a highly heterogeneous population of scrambled isomers. The replacement of the Cys residues by Ala probably permits the formation of a non-polar cavity in the LCI hydrophobic core. The solvation of this type of cavities is very unfavorable and causes a remarkable loss of enthalpy in the folded state.[48] Accordingly, although FoldX underestimates LCI_1 destabilization, it predicts a higher solvation penalty for this particular mutant. Also, this disulfide is the only one connecting the β1 and β3 strands, while the β4 and β5 strands are connected by three and two disulfide bonds, respectively. Theoretical calculations[34] suggest that in LCI the hierarchy of disulfide-bond formation critically depends on the establishment of a folding nucleus around β1 and the beginning of β2. Apparently, this folding nucleus cannot form in the absence of the Cys11–Cys34 disulfide bond, most likely because the 17 N-terminal residues of β1 and β2 strands become either unrestricted or trapped in a non-native conformation in these conditions.

Oxidative folding implies two major chemical reactions: oxidation (disulfide formation) and isomerization (disulfide reshuffling). Under typical oxidative folding conditions (Control +), the disulfide-bonded species interconvert rapidly within each disulfide ensemble by reshuffling; in contrast, oxidation between disulfide ensembles is relatively slow. Each disulfide ensemble (1S, 2S, 3S) can be considered therefore to be in thermodynamic equilibrium, and the ratios of disulfide forms in each ensemble should reflect their relative conformational stability.[49] In this situation, in addition to loop entropy, non-covalent interactions, either native or non-native, determine the distribution of disulfide species within a particular ensemble. In contrast, in the absence of free thiols (Control −), reshuffling reactions are highly restricted and an entropically determined quasi-stochastic distribution of disulfides species is expected. Comparison of disulfide profiles of single LCI mutants in the presence and in the absence of free thiols during the formation of the 1S and 2S ensembles indicates that the distribution of species does not differ significantly in both conditions for each particular mutant. This implies that non-covalent interactions do not play an important role in the early stages of LCI folding, in good agreement with the prediction that R, 1S and 2S species do not contain an extensive regular structure.

The present folding analysis of single LCI mutants demonstrates that scrambled isomers with at least two non-native disulfide bonds can effectively populate their folding pathways and in certain cases accumulate significantly at the end of the reaction. These scrambled isomers are productive species since their disulfides can rearrange to render the natively connected LCI mutants. This ability is evident in the case of LCI_2 and LCI_4 folding, where the presence of thiol agents allowing the reshuffling of these trapped species in the last stage of folding clearly increases the yields of natively connected protein relative to the Control − reaction. This effect is also observed, although less relevant, in LCI_3 folding. As discussed above, in the presence of thiol agents, the 3S disulfide ensemble can be considered to be in thermodynamic equilibrium at the end of the reaction and the ratios of native to scrambled isomers should reflect their relative conformational stability. The native disulfide protection and higher conformational stability of LCI_3 relative to any other 3S isomer are likely to be the causes of its high accumulation at equilibrium, whereas the less stable and more flexible LCI_2 and LCI_4 mutants coexist with scrambled forms. In principle, the 3S scrambled isomers would form directly by oxidation of the apparently non-structured and entropically distributed 2S ensemble and their accumulation would depend on the stability of the natively bonded forms that originate from their reshuffling. Accordingly,

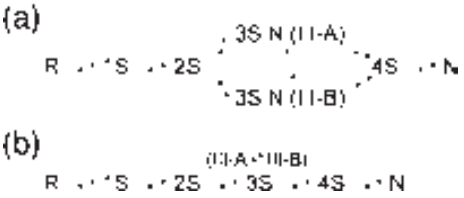


Fig. 11. Schematic diagram of the oxidative folding of LCI. (a) Scheme of the folding pathway of LCI as previously described.[33] "R" and "N" indicate the fully reduced/unfolded and native forms, respectively. 1S, 2S, 3S and 4S are ensembles of molecules with the corresponding number of disulfide bonds. "3S N" stands for 3-disulfide species with native bonds. III-A and III-B are formed by direct oxidation of 2S species already containing two native disulfide bonds. (b) Scheme of the folding pathway of LCI based on the results derived from the present work. III-A and III-B arise from a non-specific oxidation of the 2S ensemble to 3S species and the subsequent reshuffling of these forms into the native-like intermediates.

when the native protein conformation is disrupted (i.e., at high denaturant concentrations), all these mutants reshuffle rapidly to 3S scrambled forms.

Importantly, the present work questions whether III-A and III-B originate, at least exclusively, from the direct oxidation of 2S molecules containing two native disulfides. Since the analysis of the initial folding stages suggests that no enthalpic bias toward the formation of native disulfides exists, only a minor fraction of the molecules in this population is expected to display the correct combination of disulfide bonds. Once demonstrated that 3S non-natively bonded species can rearrange to render 3S forms with native disulfides and native-like conformations, a more probable mechanism for the formation of III-A and III-B intermediates would be the non-specific oxidation of the unfolded 2S ensemble to 3S species and the subsequent reshuffling of these forms into the intermediates (see Fig. 11). Thus, the folding of LCI appears to follow the quasi-stochastic model proposed by Saito *et al.* for RNase A.[50] Similar to the case of LCI_3, the high stability and/or protection of the native disulfide bonds in III-A and III-B relative to any other 3S isomer provides an explanation about why other des species cannot accumulate significantly.

III-A and III-B are "disulfide-insecure" intermediates in which the free thiol groups are protected from the solvent by the secondary structure elements; therefore, structural fluctuations that expose the thiol groups also expose the disulfide bonds, promoting the reshuffling and formation of 4S scrambled isomers instead of direct oxidation of the free thiols to obtain the native disulfide pairing.[14] The conversion of scrambled species to native protein constitutes the rate-limiting step in LCI folding. The new LCI folding scheme might imply that thiol agents accelerate the LCI folding pathway by promoting reshuffling not only in the 4S ensemble but also in the 3S population. This would be consistent with the increased rate and efficiency of the LCI folding reactions performed under mild denaturing conditions. This effect had been previously attributed to partial unfolding of III-A and III-B intermediates.[33] However, the new data indicate that folded LCI mutants displaying three native disulfides, with stabilities that overall should resemble those of natural intermediates, form efficiently in the presence of moderate concentrations of chaotropic agents. In fact, the folding pathways of LCI_2 and especially of LCI_3 are significantly accelerated, indicating a destabilization of the kinetic traps in the folding reaction of these mutants. Because the R, 1S and 2S forms are unstructured, it is unlikely that the denaturant agent exerts its action over these species; it rather destabilizes the more compact 3S non-native species, making their disulfides more accessible and promoting their reshuffling. The effect of the denaturant depends on the energy gap between the folded and the kinetically trapped species. In this way, the acceleration of the folding rate is clearly higher for LCI_3 than for LCI_2. No effect is expected when the denaturant affects to the same extent the structures of the folded and scrambled forms, as it occurs in the case of the marginally stable LCI_4 mutant.

Contribution of disulfide bonds to LCI function

Although proteins possess stable native structures, the evolutionary fitness of a protein does not depend on the stability of the native structure per se but rather on the stability of the structure that allows the protein to better perform its function in its physiological environment. Stability is therefore under selection only insofar as it is necessary for the biochemical function, and many proteins are only marginally stable at their physiologically relevant temperatures.[51] Surprisingly, LCI_2 and LCI_3 are excellent carboxypeptidase inhibitors in spite of being significantly destabilized, especially in the case of LCI_2. As previously shown for LCI_3, docking of the mutant on top of the active site of the carboxypeptidase strongly restricts the conformational fluctuation of the inhibitor and allows an efficient binding to the target enzyme.[35] This induction of the inhibitory activity upon docking to the protease has also been shown for highly destabilized forms of BPTI.[52] From the results reported here, it appears that the stability and inhibitory activity of LCI are uncoupled, since the WT stability clearly exceeds that necessary for an efficient inhibition of carboxypeptidases. However, this extra stability could allow LCI to be functional in its physiological environment. LCI is a component of leech saliva that acts in the host blood, two fluids where low stability usually results in reduced half-life due to their high content of proteolytic enzymes. In addition, according to the three-dimensional structures of III-A and LCI_3, the entropic cost of binding is higher in these forms than in the WT, whereas the enthalpic gain should be very similar. This predicts a higher dissociation constant for the 3-disulfide mutants compared with WT LCI, a fact that is indeed confirmed by inhibitory assays with LCI_3. This mutant would be a temporary inhibitor, as previously described for mutants of the *Streptomyces* subtilisin inhibitor lacking one disulfide bond that are more susceptible to proteolysis.[53] In any case, the uncoupling of thermodynamic stability and inhibitory activity in LCI may pave the way for the design of novel LCI variants with improved properties. In this context, it is worth remembering the increasing biomedical interest of both natural and synthetic carboxypeptidase inhibitors for coagulation–fibrinolytic disorders.[30,54]

Materials and Methods

Site-directed mutagenesis, protein expression and purification

The synthetic gene encoding for LCI and cloned into the pBAT-4-OmpA vector[55] was used as template for PCR in the site-directed mutagenesis. The seven mutants designed for the present work were generated using a

QuickChange Site-Directed Mutagenesis Kit (Stratagene) according to the instructions of the manufacturer and verified by DNA sequencing. The WT and mutant forms of LCI were produced and purified as previously published.[28] Briefly, the proteins were expressed in *E. coli* strain BL21(DE3), growing the cells in M9 minimal medium (0.5% glycerol) supplemented with 0.2% casamino acids at 37 °C and inducing at an A_{600} of 1.0 with 1 mM isopropyl-β-D-thiogalactopyranoside (final concentration). The secreted proteins were treated with 2 mM cystine and 4 mM cysteine (final concentrations) at pH 8.5 to obtain the maximum amount of native form. The proteins were subsequently purified from the culture supernatant using a Sep-Pak C18 cartridge (Waters), followed by anion-exchange chromatography on a TSK-DEAE 5PW column (Tosohaas) and by RP-HPLC (Alliance, Waters) on a 4.6-mm Protein C4 column (Vydac Grace). Protein identity and purity were confirmed by MALDI-TOF MS and automated Edman degradation. The concentration of the LCI variants in solution was determined by measuring the absorbance at 280 nm and using the calculated absorption coefficient $E_{0.1\%} = 2.12$. The purified and quantified proteins were kept lyophilized.

Oxidative folding experiments

The native protein (2 mg) was reduced and unfolded in 0.5 M Tris–HCl buffer, pH 8.5, containing 200 mM DTT and 6 M guanidine thiocyanate for 2 h at room temperature. To initiate folding, we loaded the sample onto a PD-10 desalting column (Sephadex G-25, GE Healthcare) previously equilibrated with 0.1 M Tris–HCl buffer, pH 8.5. The protein was eluted in 1.2 ml and split in three parts that were immediately diluted in the same buffer to a final protein concentration of 0.5 mg/ml, in the absence (Control −) and in the presence (Control +) of 0.25 mM 2-mercaptoethanol or 0.5 mM/1 mM GSSG/GSH. In some cases, the refolding experiments were carried out in the presence of 0.5 M GdnHCl. Folding intermediates were trapped in a time-course manner either by acidification with 4% aqueous trifluoroacetic acid (TFA) or by alkylation with 0.1 M Tris–HCl buffer, pH 8.5, containing 0.1 M 4-vinylpyridine for 1 h at room temperature. The acid-trapped intermediates were subsequently analyzed by RP-HPLC using a linear 20%–40% gradient of acetonitrile with 0.1% TFA over 50 min on a 4.6-mm Jupiter C4 column (Phenomenex) at a flow rate of 0.75 ml/min. Aliquots derivatized with 4-vinylpyridine (each pair of modified cysteines increases the molecular mass by 212 Da) were diluted 1:10 in 0.1% aqueous TFA and analyzed by MALDI-TOF MS using an Ultraflex spectrometer (Bruker) operating in linear mode under 20 kV. Samples were prepared by mixing equal volumes of the protein solution and matrix solution (10 mg/ml of sinapic acid dissolved in aqueous 30% acetonitrile with 0.1% TFA) and using the dried droplet method. A mixture of proteins from Bruker (protein calibration standard I; mass range = 3–25 kDa) was used as external mass calibration standard.

Reductive unfolding and disulfide scrambling experiments

The native protein (0.5 mg/ml) was dissolved at room temperature in 0.1 M Tris–HCl buffer, pH 8.5, containing increasing concentrations of DTT (1–100 mM). Aliquots of the mixtures were removed at various time points, trapped by acidification with 4% aqueous TFA and analyzed by RP-HPLC as detailed in Oxidative folding experiments to monitor the unfolding reaction. For disulfide-scrambling experiments, the native protein (0.5 mg/ml) was dissolved in 0.1 M Tris–HCl buffer, pH 8.5, containing 0.25 mM 2-mercaptoethanol as thiol initiator and selected concentrations of denaturants (0–8 M urea or GdnHCl). The reaction was allowed to reach equilibrium typically for 20 h at room temperature. The samples were then quenched with 4% aqueous TFA and analyzed by RP-HPLC as described above. Least-squares fitting of reductive unfolding data to a first-order reaction and determinations of standard errors of the fitted parameters were conducted using the program KaleidaGraph (Synergy Software, Reading, PA).

Thermodynamic parameters

As previously reported for WT LCI,[35] thermodynamic parameters were calculated by fitting experimental data from disulfide scrambling to Eq. (1):

$$y = \frac{(a + b[\text{denaturant}]) - (c + d[\text{denaturant}])}{1 + \exp\left\{-\left(\Delta G_U^{H_2O} - m[\text{denaturant}]\right)/RT\right\}} + (c + d[\text{denaturant}])$$

where y is the observed fractional population of native species, $\Delta G_U^{H_2O}$ is the free energy change of unfolding in the absence of denaturant, m is a parameter for cooperativity of unfolding and $(a + b[\text{denaturant}])$ and $(c + d[\text{denaturant}])$ are terms for the baseline dependence on denaturant concentration. The midpoint concentration of unfolding, c_M, was calculated by Eq. (2):

$$c_M = \Delta G_U^{H_2O}/m$$

The fitting was performed using the non-linear least-squares algorithm provided with the software KaleidaGraph assuming a two-state unfolding mechanism.

CD, fluorescence and NMR spectroscopy

Samples for CD spectroscopy were prepared by dissolving the protein to a final concentration of 0.2 mg/ml in 20 mM Tris–HCl buffer, pH 7.5. CD spectra were collected in a Jasco J-715 spectrometer at 25 °C using a cell of 2-mm path length. Samples for fluorescence spectroscopy were measured in the same buffer on a fluorescence spectrophotometer (Cary Eclipse, Varian) at a final protein concentration of 13 μM and at 25 °C. The excitation wavelength was set at 268 and 280 nm, with excitation and emission slit widths of 5 and 10 nm, respectively. The fluorescence emission was measured between 280 and 400 nm. Samples for NMR spectroscopy were prepared at a protein concentration of 1.5 mg/ml in 20 mM Tris–HCl buffer, pH 7.5, using a 9:1 ratio by volume of H_2O/D_2O. One-dimensional NMR spectra were acquired at 25 °C on a Bruker AVANCE 600-MHz spectrometer using solvent-suppression WATERGATE techniques. The collected spectra were processed and analyzed using the TopSpinv2.0 software packages from Bruker Biospin.

bis-ANS binding assay

The fluorescence emission spectra of bis-ANS in the absence and in the presence of protein were recorded using 20 mM Tris–HCl buffer, pH 7.5, containing 5 μM

bis-ANS to obtain a final protein concentration of 13 μM. Samples were read after 10 min of incubation using an excitation wavelength of 370 nm. The fluorescence emission was measured between 400 and 600 nm, with excitation and emission slit widths of 10 nm each. Thermal transition curves were obtained at a heating rate of 1 °C/min by measuring the bis-ANS emission at 460 nm after excitation at 370 nm.

CPA inhibitory experiments

The inhibitory activity of the native form of different LCI variants was tested by measuring the inhibition of the hydrolysis of the chromogenic substrate *N*-(4-methoxyphenylazoformyl)-Phe-OH (Bachem) by bovine CPA (Sigma) at 350 nm using a Cary 400 UV-Vis spectrophotometer (Varian). The assays were performed in 50 mM Tris–HCl buffer, pH 7.5, containing 100 mM NaCl, at a substrate concentration of 100 μM. K_i values for the complexes of WT and LCI mutants with CPA were determined using pre-steady-state kinetics as reported for tight-binding inhibitors, with a final correction for substrate competition.[56]

FoldX modeling

To model the mutations in LCI, we used the BuildModel option of the FoldX algorithm (version 2.65).[44] This command reads the Protein Data Bank coordinates and duplicates them internally. Then, it mutates the selected position in one molecule to itself and, in the other, to the selected variant, while moving the neighboring side chains. It ensures that the moving side chains and the rotamer set for them are the same in both cases. The effect of the mutation is subsequently computed by subtracting the energy of the self-mutated WT from that of the mutant. The $\Delta\Delta G$ values are provided in kilocalories per mole for all modeled structures.

Acknowledgements

This study was supported by grant BIO2007-68046 from the Spanish Ministry of Science and Innovation and grants SGR01037 and SGR00037 from the national Catalan government. V.C. is a beneficiary of a predoctoral fellowship from the Spanish Ministry of Science and Innovation. We acknowledge Prof. Christian P. Sommerhoff and two anonymous reviewers for their helpful suggestions to the manuscript.

References

1. Creighton, T. E., Zapun, A. & Darby, N. J. (1995). Mechanisms and catalysts of disulfide bond formation in proteins. *Trends Biotechnol.* **13**, 18–23.
2. Frand, A. R., Cuozzo, J. W. & Kaiser, C. A. (2000). Pathways for protein disulphide bond formation. *Trends Cell Biol.* **10**, 203–210.
3. Hogg, P. J. (2003). Disulfide bonds as switches for protein function. *Trends Biochem. Sci.* **28**, 210–214.
4. Kawamura, S., Ohkuma, M., Chijiiwa, Y., Kohno, D., Nakagawa, H., Hirakawa, H. *et al.* (2008). Role of disulfide bonds in goose-type lysozyme. *FEBS J.* **275**, 2818–2830.
5. Klink, T. A., Woycechowsky, K. J., Taylor, K. M. & Raines, R. T. (2000). Contribution of disulfide bonds to the conformational stability and catalytic activity of ribonuclease A. *Eur. J. Biochem.* **267**, 566–572.
6. Liu, Y., Breslauer, K. & Anderson, S. (1997). "Designing out" disulfide bonds: thermodynamic properties of 30–51 cystine substitution mutants of bovine pancreatic trypsin inhibitor. *Biochemistry*, **36**, 5323–5335.
7. Vaz, D. C., Rodrigues, J. R., Sebald, W., Dobson, C. M. & Brito, R. M. (2006). Enthalpic and entropic contributions mediate the role of disulfide bonds on the conformational stability of interleukin-4. *Protein Sci.* **15**, 33–44.
8. Zavodszky, M., Chen, C. W., Huang, J. K., Zolkiewski, M., Wen, L. & Krishnamoorthi, R. (2001). Disulfide bond effects on protein stability: designed variants of *Cucurbita maxima* trypsin inhibitor-V. *Protein Sci.* **10**, 149–160.
9. Kurokawa, Y., Koganesawa, N., Kobashigawa, Y., Koshiba, T., Demura, M. & Niita, K. (2001). Oxidative folding of human lysozyme: effects of the loss of two disulfide bonds and the introduction of a calcium-binding site. *J. Protein Chem.* **20**, 293–303.
10. Parrini, C., Bemporad, F., Baroncelli, A., Gianni, S., Travaglini-Allocatelli, C., Kohn, J. E. *et al.* (2008). The folding process of acylphosphatase from *Escherichia coli* is remarkably accelerated by the presence of a disulfide bond. *J. Mol. Biol.* **379**, 1107–1118.
11. Qin, M., Zhang, J. & Wang, W. (2006). Effects of disulfide bonds on folding behavior and mechanism of the beta-sheet protein tendamistat. *Biophys. J.* **90**, 272–286.
12. Ruoppolo, M., Vinci, F., Klink, T. A., Raines, R. T. & Marino, G. (2000). Contribution of individual disulfide bonds to the oxidative folding of ribonuclease A. *Biochemistry*, **39**, 12033–12042.
13. Wedemeyer, W. J., Welker, E., Narayan, M. & Scheraga, H. A. (2000). Disulfide bonds and protein folding. *Biochemistry*, **39**, 4207–4216.
14. Welker, E., Narayan, M., Wedemeyer, W. J. & Scheraga, H. A. (2001). Structural determinants of oxidative folding in proteins. *Proc. Natl Acad. Sci. USA*, **98**, 2312–2316.
15. Creighton, T. E. (1986). Disulfide bonds as probes of protein folding pathways. *Methods Enzymol.* **131**, 83–106.
16. Creighton, T. E. (1997). Protein folding coupled to disulphide bond formation. *Biol. Chem.* **378**, 731–744.
17. Narayan, M., Welker, E., Wedemeyer, W. J. & Scheraga, H. A. (2000). Oxidative folding of proteins. *Acc. Chem. Res.* **33**, 805–812.
18. Chang, J. Y. (2004). Evidence for the underlying cause of diversity of the disulfide folding pathway. *Biochemistry*, **43**, 4522–4529.
19. Arolas, J. L., Aviles, F. X., Chang, J. Y. & Ventura, S. (2006). Folding of small disulfide-rich proteins: clarifying the puzzle. *Trends Biochem. Sci.* **31**, 292–301.
20. Chang, J. Y. (2008). Diversity of folding pathways and folding models of disulfide proteins. *Antioxid. Redox Signal.* **10**, 171–178.
21. Darby, N. J., Morin, P. E., Talbo, G. & Creighton, T. E. (1995). Refolding of bovine pancreatic trypsin inhibitor via non-native disulphide intermediates. *J. Mol. Biol.* **249**, 463–477.
22. Weissman, J. S. & Kim, P. S. (1991). Reexamination of the folding of BPTI: predominance of native intermediates. *Science*, **253**, 1386–1393.

23. Chang, J. Y. (1996). The disulfide folding pathway of tick anticoagulant peptide (TAP), a Kunitz-type inhibitor structurally homologous to BPTI. *Biochemistry*, **35**, 11702–11709.
24. Chang, J. Y., Canals, F., Schindler, P., Querol, E. & Aviles, F. X. (1994). The disulfide folding pathway of potato carboxypeptidase inhibitor. *J. Biol. Chem.* **269**, 22087–22094.
25. Chatrenet, B. & Chang, J. Y. (1993). The disulfide folding pathway of hirudin elucidated by stop/go folding experiments. *J. Biol. Chem.* **268**, 20988–20996.
26. Chang, J. Y., Li, L. & Lai, P. H. (2001). A major kinetic trap for the oxidative folding of human epidermal growth factor. *J. Biol. Chem.* **276**, 4845–4852.
27. Salamanca, S., Li, L., Vendrell, J., Aviles, F. X. & Chang, J. Y. (2003). Major kinetic traps for the oxidative folding of leech carboxypeptidase inhibitor. *Biochemistry*, **42**, 6754–6761.
28. Reverter, D., Vendrell, J., Canals, F., Horstmann, J., Aviles, F. X., Fritz, H. & Sommerhoff, C. P. (1998). A carboxypeptidase inhibitor from the medical leech *Hirudo medicinalis*. Isolation, sequence analysis, cDNA cloning, recombinant expression, and characterization. *J. Biol. Chem.* **273**, 32927–32933.
29. Reverter, D., Fernandez-Catalan, C., Baumgartner, R., Pfander, R., Huber, R., Bode, W. *et al.* (2000). Structure of a novel leech carboxypeptidase inhibitor determined free in solution and in complex with human carboxypeptidase A2. *Nat. Struct. Biol.* **7**, 322–328.
30. Arolas, J. L., Vendrell, J., Aviles, F. X. & Fricker, L. D. (2007). Metallocarboxypeptidases: emerging drug targets in biomedicine. *Curr. Pharm. Des.* **13**, 349–366.
31. Marx, P. F. (2004). Thrombin-activatable fibrinolysis inhibitor. *Curr. Med. Chem.* **11**, 2335–2348.
32. Salamanca, S., Villegas, V., Vendrell, J., Li, L., Aviles, F. X. & Chang, J. Y. (2002). The unfolding pathway of leech carboxypeptidase inhibitor. *J. Biol. Chem.* **277**, 17538–17543.
33. Arolas, J. L., Bronsoms, S., Lorenzo, J., Aviles, F. X., Chang, J. Y. & Ventura, S. (2004). Role of kinetic intermediates in the folding of leech carboxypeptidase inhibitor. *J. Biol. Chem.* **279**, 37261–37270.
34. Arolas, J. L., D'Silva, L., Popowicz, G. M., Aviles, F. X., Holak, T. A. & Ventura, S. (2005). NMR structural characterization and computational predictions of the major intermediate in oxidative folding of leech carboxypeptidase inhibitor. *Structure (Camb.)*, **13**, 1193–1202.
35. Arolas, J. L., Popowicz, G. M., Bronsoms, S., Aviles, F. X., Huber, R., Holak, T. A. & Ventura, S. (2005). Study of a major intermediate in the oxidative folding of leech carboxypeptidase inhibitor: contribution of the fourth disulfide bond. *J. Mol. Biol.* **352**, 961–975.
36. Chang, J. Y., Li, L. & Bulychev, A. (2000). The underlying mechanism for the diversity of disulfide folding pathways. *J. Biol. Chem.* **275**, 8287–8289.
37. Li, Y. J., Rothwarf, D. M. & Scheraga, H. A. (1995). Mechanism of reductive protein unfolding. *Nat. Struct. Biol.* **2**, 489–494.
38. Chang, J. Y. (1997). A two-stage mechanism for the reductive unfolding of disulfide-containing proteins. *J. Biol. Chem.* **272**, 69–75.
39. Chang, J. Y. & Li, L. (2001). The structure of denatured alpha-lactalbumin elucidated by the technique of disulfide scrambling: fractionation of conformational isomers of alpha-lactalbumin. *J. Biol. Chem.* **276**, 9705–9712.
40. Takashi, R., Tonomura, Y. & Morales, M. F. (1977). 4,4′-Bis (1-anilinonaphthalene 8-sulfonate) (bis-ANS): a new probe of the active site of myosin. *Proc. Natl Acad. Sci. USA*, **74**, 2334–2338.
41. Pace, C. N., Grimsley, G. R., Thomson, J. A. & Barnett, B. J. (1988). Conformational stability and activity of ribonuclease T1 with zero, one, and two intact disulfide bonds. *J. Biol. Chem.* **263**, 11820–11825.
42. Wang, M. C. & Uhlenbeck, G. E. (1945). On the theory of the Brownian motion II. *Rev. Mod. Phys.* **17**, 323–342.
43. Guerois, R., Nielsen, J. E. & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* **320**, 369–387.
44. Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F. & Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Res.* **33**, W382–W388.
45. Darby, N. & Creighton, T. E. (1995). Disulfide bonds in protein folding and stability. *Methods Mol. Biol.* **40**, 219–252.
46. Ewbank, J. J. & Creighton, T. E. (1993). Pathway of disulfide-coupled unfolding and refolding of bovine alpha-lactalbumin. *Biochemistry*, **32**, 3677–3693.
47. Serrano, L., Kellis, J. T., Jr, Cann, P., Matouschek, A. & Fersht, A. R. (1992). The folding of an enzyme: II. Substructure of barnase and the contribution of different interactions to protein stability. *J. Mol. Biol.* **224**, 783–804.
48. Hubbard, S. J. & Argos, P. (1995). Evidence on close packing and cavities in proteins. *Curr. Opin. Biotechnol.* **6**, 375–381.
49. Welker, E., Wedemeyer, W. J., Narayan, M. & Scheraga, H. A. (2001). Coupling of conformational folding and disulfide-bond reactions in oxidative folding of proteins. *Biochemistry*, **40**, 9059–9064.
50. Saito, K., Welker, E. & Scheraga, H. A. (2001). Folding of a disulfide-bonded protein species with free thiol(s): competition between conformational folding and disulfide reshuffling in an intermediate of bovine pancreatic ribonuclease A. *Biochemistry*, **40**, 15002–15008.
51. Fersht, A. (1999). Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding Freeman, New York, NY.
52. Krokoszynska, I., Dadlez, M. & Otlewski, J. (1998). Structure of single-disulfide variants of bovine pancreatic trypsin inhibitor (BPTI) as probed by their binding to bovine beta-trypsin. *J. Mol. Biol.* **275**, 503–513.
53. Kojima, S., Kumagai, I. & Miura, K. (1993). Requirement for a disulfide bridge near the reactive site of protease inhibitor SSI (*Streptomyces* subtilisin inhibitor) for its inhibitory action. *J. Mol. Biol.* **230**, 395–399.
54. Sanglas, L., Valnickova, Z., Arolas, J. L., Pallares, I., Guevara, T., Sola, M. *et al.* (2008). Structure of activated thrombin-activatable fibrinolysis inhibitor, a molecular link between coagulation and fibrinolysis. *Mol. Cell*, **31**, 598–606.
55. Peranen, J., Rikkonen, M., Hyvonen, M. & Kaariainen, L. (1996). T7 vectors with modified T7lac promoter for expression of proteins in *Escherichia coli*. *Anal. Biochem.* **236**, 371–373.
56. Bieth, J. G. (1995). Theoretical and practical aspects of proteinase inhibition kinetics. *Methods Enzymol.* **248**, 59–84.

RESEARCH ARTICLE

Deciphering the role of the thermodynamic and kinetic stabilities of SH3 domains on their aggregation inside bacteria

Virginia Castillo, Alba Espargaró, Veronica Gordo, Josep Vendrell and Salvador Ventura

Institut de Biotecnologia i de Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain

Received: April 20, 2010
Revised: June 23, 2010
Accepted: July 16, 2010

The formation of insoluble deposits by globular proteins underlies the onset of many human diseases. Recent studies suggest a relationship between the thermodynamic stability of proteins and their *in vivo* aggregation. However, it has been argued that, in the cell, the occurrence of irreversible aggregation might shift the system from equilibrium, in such a way that it could be the rate of unfolding and associated kinetic stability instead of the conformational stability that controls protein deposition. This is an important but difficult to decipher question, because kinetic and thermodynamic stabilities appear usually correlated. Here we address this issue by comparing the *in vitro* folding kinetics and stability features of a set of non-natural SH3 domains with their aggregation properties when expressed in bacteria. In addition, we compare the *in vitro* stability of the isolated domains with their effective stability in conditions that mimic the cytosolic environment. Overall, the data argue in favor of a thermodynamic rather than a kinetic control of the intracellular aggregation propensities of small globular proteins in which folding and unfolding velocities largely exceed aggregation rates. These results have implications regarding the evolution of proteins.

Keywords:
Protein aggregation / Protein folding / Protein stability / Systems Biology

1 Introduction

Protein misfolding and aggregation into insoluble amyloid fibrils or intracellular inclusions is linked to the onset of many degenerative disorders, ranging from Alzheimer's disease to Diabetes [1–4]. The sequential and conformational determinants of the assembly of isolated proteins into amyloid structures have been extensively investigated *in vitro* [5–8]. However, much less is known about the traits governing the aggregation of polypeptides into amyloid-like conformations in the highly complex cellular background. This is mainly due to the lack of simple but physiologically relevant systems in which to model *in vivo* protein aggregation. One such system could be the inclusion bodies (IBs) formed in bacteria during homologous or heterologous protein expression. Increasing evidence suggesting that they contain amyloid-like structures make of *Escherichia coli* a suitable organism in which to study the molecular determinants of protein aggregation and more specifically of amyloid formation in the cell [9–13].

The proteins and peptides responsible for amyloid diseases are not related in terms of sequence or structure. Some of these polypeptides are devoid of any regular secondary structure and remain mostly unstructured in solution [14]. Other amyloidogenic proteins are globular in their native state, implying that a properly packed and cooperatively sustained structure dominates the conformational ensemble under physiological conditions [15]. The molecular determinants of the aggregation of both types of

Correspondence: Dr. Salvador Ventura, Institut de Biotecnologia i de Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain
E-mail: salvador.ventura@uab.es
Fax: +34-935811264

Abbreviations: IBs, inclusion bodies; **SPC-SH3,** Spectrin-SH3; **WT,** wild-type

proteins in cellular environments are beginning to be studied using *E. coli* as a model system. In this way, Chiti and co-workers have recently demonstrated that, for fully or partially unfolded proteins, intrinsic protein sequential properties shown to modulate amyloid fibril formation *in vitro* [16] like hydrophobicity, β-sheet propensity and charge also govern protein aggregation inside bacteria [17]. For globular proteins, the study of three different protein models suggest that the conformational stability of the folded native state may act as a major determinant of protein deposition in physiological environments [18–20]. In particular, in a recent study, we have used the Spectrin-SH3 (SPC-SH3) domain as a model of small globular domain to study the aggregation properties of the wild-type (WT) protein and mutants thereof inside bacteria [20]. We found that, as a trend, destabilized mutants tend to be more aggregation-prone than the WT and that stabilization of the native fold increases solubility. During the revision of that work, one of the anonymous reviewers put forward an important question: Whether the apparent correlation we and others have observed between protein aggregation and stability in physiological backgrounds would reflect the effect of the energy barriers to unfolding or kinetic stability rather than thermodynamic stability itself. This question is pertinent, because the effects of mutations on the thermodynamic and kinetic stability of proteins are often related [21]. According to protein engineering analysis [22], mutations affecting the stability of the native state and occurring in regions of the protein that become unstructured in the transition state would have a ϕ value of 0 and are expected to exhibit parallel effects on the thermodynamic and kinetic stability. In contrast, mutations at residues ordered in the transition state ensemble would render a ϕ value of 1 and only affect the thermodynamic stability. Because most protein mutations result in intermediate ϕ values, both parameters become usually associated, and deciphering their individual contribution to a given phenomenon and in particular to aggregation is not a trivial task. Importantly, there is a reduced, but increasing, number of globular proteins as transthyretin [23], T7 endonuclease I [24], ataxin-3 [25], prion protein [26] and Z α1-antitrypsin [27] in which aggregation promoting mutations do not correlate with a decrease in thermodynamic stability but rather with changes in the unfolding rates.

At that time, we could not answer the proposed question because we only had qualitative data on the aggregation properties of the different domains and lacked information on the folding kinetics of many of the proteins in the study. In fact, in spite of being a crucial issue to understand how evolution shapes protein properties, the relationship between folding and unfolding rates and the aggregation properties of proteins in living organisms remains essentially unexplored from the experimental point of view. Therefore, in the present work we have made an effort to quantify the solubility of the collection of SH3 domains inside bacteria, to calculate their *in vitro* folding kinetic properties and to approximate the effective conformational stabilities of these proteins in cellular extracts in order to contribute to provide a quantitative and comprehensive view of the physical principles underlying the aggregation of small globular proteins inside the cell.

2 Materials and methods

2.1 Cloning

PCR was used to build 15 mutant DNA sequences containing from three to nine amino acid substitutions in the folding core and three additional single mutants, as previously described [28, 29]. The PCR products were cloned into vector pBAT-4 and the constructs were used to transform XL-1 Blue *E. coli* cells. Chemical sequencing of the vectors purified from positive clones confirmed the individual sequences of the 18-variant set used for the experiments. The same protocol was used to generate a truncated SPC-SH3 mutant lacking the five C-terminal residues [30].

2.2 Protein expression, extraction and purification

Purified recombinant pBAT-4 expression vectors were used to transform *E. coli* BL21 (DE3) cells by the heat shock method. Cells were grown overnight from a single colony in Luria-Bertani (LB) medium containing 50 mg/L ampicillin. An aliquot was subsequently inoculated into fresh medium at a 1/100 dilution to reach a final volume of 1 L. When the OD_{600} of the culture reached values between 0.4 and 0.6, the expression of the recombinant variants was induced by the addition of IPTG to a final concentration of 1 mM. Three hours after the induction time, cells were harvested by centrifugation at 5000 rpm for 20 min, resuspended in 10 mL of PBS containing *Pefabloc* (0.4 mg/mL), lysed by four consecutive freeze–thawing cycles followed by three 30 s sonications and finally ultracentrifuged at 16 000 rpm for 30 min. Pellets and supernatants were analyzed by SDS-PAGE to determine which fractions contained the recombinant proteins. Proteins from the insoluble fraction were resuspended in 6 M urea, 10 mM sodium citrate, 100 mM NaCl (pH 3.5), ultracentrifuged at 16 000 rpm for 30 min and loaded onto a a HiLoad 26/60 Superdex 75 column equilibrated with the same buffer. The fractions contaning pure protein were pooled together, diluted ten times in the same buffer devoid of urea, concentrated by ultrafiltration, dialysed against water and kept frozen until their analysis. The fraction containing soluble proteins was supplemented with polyethyleneimime to precipitate DNA fragments. The supernatants recovered after a further ultracentrifugation at 16 000 rpm for 30 min were treated with ammonium sulfate in two steps; in the first one, the salt was added to 30% concentration to precipitate high molecular weight proteins that were separated by ultracentrifugation at 16 000 rpm for

30 min. A further addition of ammonium sulfate to the supernatant until reaching a 70% concentration allowed for the precipitation of the SH3 domains, which were recovered after a final ultracentrifugation at 16 000 rpm for 45 min. The pellets were redissolved in 10 mM sodium citrate, 100 mM NaCl (pH 3.5) dialysed against water and kept frozen until their analysis.

2.3 Folding kinetic experiments

The kinetics of the folding and unfolding reactions were followed in a Bio-Logic stopped-flow fluorimeter using excitation and measuring wavelengths of 290 and 320 nm, respectively. Starting with the protein in 50 mM sodium phosphate buffer (pH 7.0), the unfolding reaction was promoted by dilution with appropriate volumes of the same buffer containing increasing concentrations of urea. Conversely, the unfolding reaction was followed by addition of appropriate volumes of urea-free buffer to an initial protein solution in 50 mM sodium phosphate buffer, 9.5 M urea (pH 7.0). The experiments were performed at 298 K. The individual folding and unfolding reactions rates were calculated for at least 15 different final urea concentrations for each protein under study.

From the observation of a curvature in the plot of the natural logarithm of the unfolding rate *versus* urea concentration when studying the unfolding kinetics of SH3, Serrano and co-workers [31] developed and equation that allows the calculation of the slopes of the curves and the rate constants of the refolding and unfolding reactions (see a description of the parameters in the legend to Table 1). From theses values it is possible to calculate the destabilization energy induced by the mutations, $\Delta\Delta G_{F\text{-}U}$, and its two semi-reaction components:

$$\Delta\Delta G_{F-U} = (\Delta G_{F-U})_{wt} - (\Delta G_{F-U})_{mut} = \Delta\Delta G_{\ddagger-U} - \Delta\Delta G_{\ddagger-F}$$

$$\Delta\Delta G_{\ddagger-U} = RT \ln (k_f^{wt}/k_f^{mut})$$

$$\Delta\Delta G_{\ddagger-F} = RT \ln (k_u^{wt}/k_u^{mut})$$

where ''wt'' refers to the SPC-SH3 domain and ''mut'' to each one of the mutations done upon the former.

2.4 Thermal denaturation of SH3 domains

A Jasco J-715 spectropolarimeter equipped with a thermostat-controlled cell holder was used to study the thermal denaturation of SH3 domains (at 20 μM concentration) following changes in the circular dichroism signal at 220 nm at temperatures ranging from 15 to 95°C and using a scan rate of 1°C/min. Two independent scans were performed for each analyzed mutant. Fittings to calculate temperature (Tm) were done using the non-linear least squares algorithm provided with Kaleidagraph (Abelbeck Software, Reading, PA).

2.5 Quantification of protein solubility

Protein solubility was measured as described in [20]. Briefly, *E. coli* BL21 (DE3) cells transformed with recombinant pBAT-4 expression vectors were grown and treated as described under Section 2.2. Cells from 1500 μL aliquots were harvested by centrifugation and resuspended in 300 μL of BugBuster Protein Extraction Reagent (Novagene) to isolate the soluble and insoluble fractions. The resuspended cells were submitted to vigorous vortexing and three freeze–thawing cycles (at −80 and 37°C). Samples were subsequently shaken at room temperature for 2 min and centrifuged at 16 000 rpm for 10 min. The supernatants and the pellets correspond to the soluble and insoluble fractions, respectively, which were analyzed by Tricine-SDS gels with 12% w/v polyacrylamide, stained with Coomassie brilliant blue and scanned at high resolution. Quantification of the electrophoretic bands was performed with the software Quantity One from Bio Rad. For each particular mutant at least three different expression assays were performed and solubility averaged. Typical variations in the soluble/insoluble ratio for a given SH3 domain are in the range 5–20% (Table 1).

2.6 Aggregation of SH3 domains in cell lysates

E. coli BL21 (DE3) cells transformed with recombinant pBAT-4 expression vectors were incubated in LB medium (with 100 μg/mL of ampicillin) overnight at 37°C and then inoculated into fresh medium at a 1/100 dilution. At an OD_{600} of 0.4, 10 mL aliquots were transferred to 18°C, induced with 1 mM IPTG when $OD_{600} = 0.6$ and incubated at 18°C for 12 h. Cells were pelleted for 20 min at 4000 rpm at 4°C, washed with sterile PBS, pelleted again and resuspended in 1 mL of 50 mM Tris-HCl, 100 mM NaCl, 1 mM EDTA (pH 7.5) containing protease inhibitory cocktail set III (Calbiochem). Cells were lysed by sonication and the lysate was centrifuged for 20 min at 14 000 rpm at 4°C and the cytosolic fraction used for further analysis. Aliquots of the soluble fraction were incubated at 18, 37 or 50°C for 12 h, centrifuged 30 min at 14 000 rpm and the remaining soluble protein resolved on Tricine-SDS/12% (w/v) PAGE gels. In time-course experiments, aliquots of the soluble cytosolic fractions incubated at 37°C were extracted at different time points, analyzed in the same manner and soluble SH3 proteins quantified by gel densitometry as above. For each particular mutant at least two different assays were performed and solubility averaged.

2.7 Stability of SH3 domains in front of chemical denaturation monitored by pulse proteolysis

Equilibrium unfolding properties of purified SH3 domains and recombinant SH3 domains in complex cell lysates were

Table 1. Intracellular aggregation and equilibrium and kinetic folding parameters of SPC-SH3 domains

Protein	Solubility[a)]	k_u[b)]	k_f[c)]	$\Delta\Delta G_{‡-F}$[d)]	$\Delta\Delta G_{‡-U}$[e)]	$\Delta\Delta G_{F-U}$[f)]	m_u[g)]	m_f[h)]	$\Delta\Delta G_{F-U}$[i)]
Best9	**17.4±2.8**	**1.377±0.1**	**3.3±0.1**	**−3.5**	**0.1**	**3.6**	**0.35±0.02**	**0.95±0.09**	**3.1**
Best2	**18.2±3.9**	**0.078±0.007**	**0.6±0.08**	**−1.8**	**1.1**	**2.9**	**1.03±0.08**	**0.87±0.10**	**2.4**
MAXL	**19.4±4.3**	**0.101±0.04**	**0.2±0.06**	**−1.9**	**1.8**	**3.7**	**0.75±0.04**	**0.80±0.05**	**2.8**
MAXF	**21.6±5.8**	**1.323±0.1**	**7.9±0.2**	**−3.4**	**−0.4**	**3.1**	**0.80±0.06**	**0.80±0.06**	**2.6**
Best7	**23.8±6.1**	**1.127±0.09**	**10.2±0.4**	**−3.3**	**−0.6**	**2.8**	**1.05±0.09**	**1.10±0.07**	**2.5**
M25A	24.2±7.7	0.270±0.03	1.7±0.05	−2.5	0.5	3	0.49±0.02	0.83±0.06	2.6
MAXI	**27.4±4.6**	**0.037±0.004**	**1.5±0.1**	**−1.3**	**0.6**	**1.9**	**0.93±0.03**	**0.87±0.06**	**1.5**
Best4	52.5±4.8	0.008±0.002	16.1±0.5	0.4	−0.8	−0.4	0.53±0.04	0.80±0.01	−0.5
Best5	53.1±6.3	0.019±0.003	22.8±0.6	−0.9	−1.0	−0.1	0.58±0.02	0.79±0.01	−0.2
SH3 WT	55.9±6.3	0.0045±0.0005	3.9±0.1	0	0	0	0.47±0.01	0.87±0.01	0
C8A	56.1±8.6	0.133±0.009	49.1±1.1	−2.1	−1.5	0.6	0.60±0.01	0.74±0.02	0.1
B4-I25V[1]	63.4±4.3	0.0010±0.0001	12.3±0.5	0.8	−0.7	−1.5	0.66±0.04	0.81±0.01	−1.5
B5-I25V[1]	65.5±5.6	0.0025±0.0008	14.7±0.5	0.4	−0.8	−1.2	0.66±0.05	0.79±0.01	−1.0
N47G	68.1±6.2	0.004±0.0002	5.1±0.3	−0.1	0.4	0.5	0.42[j)]	0.89±0.02	−0.4
D48G	70.3±7.4	0.011±0.001	77±5	−0.6	−1.8	−1.2	0.43[j)]	0.84±0.02	−1.7
C8A-I25V	71.6±5.6	0.004±0.001	32.6±0.8	0.0	−1.3	−1.2	0.64±0.03	0.76±0.01	−1.7
B4-I44V[1]	79.7±9.1	0.0010±0.0005	21.8±0.6	0.8	−1.0	−1.8	0.59±0.04	0.79±0.01	−2.2
B5-I44V[1]	81.2±9.8	0.0028±0.0008	32.7±0.5	0.2	−1.3	−1.5	0.65±0.02	0.78±0.01	−2.0
C8A-I53V	85.8±6.1	0.0028±0.0008	32.7±0.5	0.2	−1.3	−1.4	0.65±0.02	0.78±0.01	−2.0

SPC-SH3 domains which folding kinetics have been experimentally determined in the present study are shown in bold. Other kinetic and equilibrium data are from [28, 29, 31].
a) Percentage of protein in the soluble fraction (average of three expression experiments).
b) Unfolding rate constant in water (s^{-1}).
c) Refolding rate constant in water (s^{-1}).
d) Difference between the WT protein and mutant unfolding activation energy (kcal/mol).
e) Difference between the mutant and WT folding activation energy (kcal/mol).
f) Gibbs energy of unfolding determined from the kinetic parameters (kcal/mol).
g) Dependence of the natural logarithm of unfolding with urea (kcal/mol M).
h) Dependence of the natural logarithm of refolding with urea (kcal/mol M).
i) Gibbs energy of unfolding measured at equilibrium (kcal/mol).
j) Value fixed in the fitting.

analyzed by pulse proteolysis in the presence of increasing concentrations of urea [32]. For purified SH3 domains a concentration of 0.4 mg/mL was used. For the soluble cytosolic fractions, the protein concentration was adjusted with buffer in such a way that the concentration of recombinant SH3 domains in the extract was ~0.4 mg/mL. The protein solutions were equilibrated overnight at 25°C in 20 mM Tris–HCl buffer (pH 8.0) containing 10 mM $CaCl_2$, 50 mM NaCl and varying concentrations of urea. The equilibrated protein solutions were heated to 37°C and treated with 0.20 mg/mL thermolysin for 2 min and then quenched with 50 mM EDTA. Urea-dependent proteolysis of SH3 domains was analyzed on Tricine-SDS/12% (w/v) PAGE gels as described. At least two independent experiments were performed for both purified proteins and cytosolic extracts to ensure reproducibility.

2.8 Stability of SH3 domains in front of proteolysis under non-denaturing conditions

Purified proteins and soluble cytosolic fractions containing the proteins of interest at ~0.4 mg/mL in 20 mM Tris–HCl buffer (pH 8.0) containing 10 mM $CaCl_2$, 50 mM NaCl were treated with 0.20 mg/mL thermolysin and incubated at 37°C for 1 h. The digestion was quenched at selected time points with 50 mM EDTA. The time-course of the reaction was analyzed on Tricine-SDS/12% (w/v) PAGE gels as described. At least two independent experiments were performed for both purified proteins and cytosolic extracts to ensure reproducibility.

3 Results and discussion

3.1 *In vivo* aggregation propensities of SH3 domains

SPC-SH3 is a small 62-residue protein domain. It consists of two β-sheets that form an orthogonal sandwich structure. The folding/unfolding of SPC-SH3 follows a two-state transition mechanism *in vitro* [33, 34]. In a previous work, the WT domain and a set of mutants were expressed in bacteria and could be classified based on their differential solubility as (i) mostly soluble upon expression, (ii) partially soluble or (iii) mainly accumulated as IBs [20]. Here, the distribution of the WT protein and 18 of the mutants between the soluble fraction and insoluble IBs after

expression has been quantified more precisely by densitometry analysis of the corresponding SPC-SH3 bands in Tricine-SDS gels in triplicate expression experiments (Table 1). The proteins display a broad range of solubilities, from mutants that accumulate as IBs in proportions higher than 80% under the conditions of the experiment (Best9, Best2 and MAXL) to mutants that are more than 80% soluble in the cytosol (C8A-I53V and B5-I44V). This dynamic range of solubilities, with nine mutants displaying lower and the other nine mutants larger solubilities than the WT domain, allowed for a complete comparative analysis of SH3 aggregation tendencies and *in vitro* thermodynamic and kinetic protein properties.

3.2 Folding kinetics of *in vivo* insoluble SH3 domains

We have previously characterized the folding kinetics of 12 of the mutants in the protein set [35]. However, all of them (except M25A) correspond to mutants with similar or higher solubilities than the WT (Table 1). Accordingly, in the present work we have focused on the aggregation-prone mutants (Best9, Best2, MAXL, MAXF, Best7 and MAXI) to have an accurate description of the folding properties associated to the complete solubility spectrum. Their kinetics of folding and unfolding were analyzed by stopped-flow at 25°C under a wide range of denaturant conditions. The folding traces obtained from fluorescence measurements fit well into single exponentials indicating the lack of detectable intermediates according with a two-state model. In all cases, a good agreement was found between the thermodynamic and kinetic values as expected for a two-state transition (Table 1). All the aggregation-prone variants unfold faster than the WT domain and display accordingly smaller unfolding activation-free energies or kinetic stabilities (Table 1). In this group, MAXI and Best9 display the lowest and highest unfolding rates, with $k_u \approx 10$- and 350-fold higher than the WT, respectively, while MAXL and Best7 display the lowest and highest refolding rates, with $k_f \approx 20$-fold lower and 3-fold higher than the WT, respectively. Besides Best7, only MAXF folds faster than the WT domain (Table 1).

Overall, in the complete 19 protein set considered in the present study unfolding and refolding rates span between two and three orders of magnitude, with proteins refolding and unfolding faster and slower than the WT. As we will see in the following sections, this broad dynamic range allowed us to dissect the contribution of the kinetic parameters to the intracellular aggregation of recombinant SH3 domains in bacteria.

3.3 Correlation between *in vitro* thermodynamic stability and *in vivo* aggregation

The conversion of globular proteins from their folded states into amyloid conformations has been extensively characterized *in vitro* using many protein models and a large variety of experimental conditions [36]. The data converge to indicate that, in general, the native structure plays a protective role against aggregation and that the first step in the conversion of normally folded proteins into fibrillar aggregates involves unfolding into fully unfolded or, more frequently, partially folded states that constitute the self-assembly competent species [1]. Accordingly, mutations or experimental conditions that destabilize the native state have been shown to result in the formation of amyloid fibrils even if this effect is only marginal or occurs transiently [37]. Actually, *in vitro* amyloid propensity and thermodynamic stability have been shown to anti-correlate for mutants of different model systems [8, 38, 39].

Three different model systems have been used to date to explore the correlation between thermodynamic stability and intracellular protein aggregation using *E. coli* as host organism, namely HypF-N, p53 and SH3 domains [18–20]. In all three cases, and in line with *in vitro* results, a qualitative negative correlation between the formation of insoluble aggregates and conformational stability was observed. However, this correlation could not be precisely quantified, either because the exact soluble/insoluble ratio could not be measured or because the mutants exhibited a reduced range of solubilities or conformational stabilities.

In Fig. 1 we plot the relationship between the measured solubility in *E. coli* and the *in vitro* determined thermodynamic stability at 25°C [35] for the 19 SH3 domains in the present study. A surprisingly high and significant linear

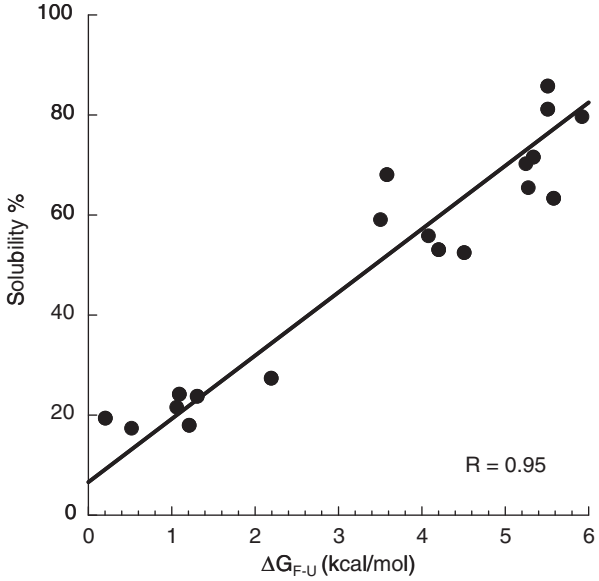


Figure 1. Correlation between the thermodynamic stability of SPC-SH3 domains and their *in vivo* solubility. Solubility is expressed as the experimentally determined percentage of SH3 domain located in the soluble cell fraction relative to the total recombinant protein. Stability was calculated at 25°C, pH 7.0, in 50 mM sodium phosphate.

correlation is observed between these two parameters ($r = 0.95$ and $p < 0.00001$), supporting the view that thermodynamic stability is one of the main parameters controlling the incorporation of SH3 domains into insoluble IBs. Under our experimental conditions, according to the equation describing the linear regression, an *in vitro* destabilization of 1 kcal/mol in the native state of an SH3 domain results in an $\approx$10% increase in its intracellular deposition propensity and *vice versa*.

3.4 Correlation between energy barriers to folding and *in vivo* aggregation

The formation of IBs in bacteria has been observed during the expression of many proteins that are unrelated in sequence, structure, size or origin [40]. For long time, this has lead to the assumption that IBs form as a result of non-specific contacts between polypeptide chains as they emerge from ribosomes and before they can attain the native protective structure [41]. During recombinant protein production the use of strong promoters, like the T7 promoter used in this work, results in high translation rates and therefore in a continuous supply of, in principle, initially unfolded polypeptides in which hydrophobic side chains and backbone hydrogen donors/acceptors might be at least transiently exposed to solvent, and therefore susceptible to establish anomalous intermolecular interactions that could lead to their aggregation into IBs. Indeed, from the y-intercept of the graph in Fig. 1 it can be expected that, under our experimental conditions, $\approx$90% of the protein would accumulate into IBs for a fully unfolded SPC-SH3 domain. This is in good agreement with the 8.5 ± 3.6 experimental solubility we have measured for MAXW, a domain with a negative Gibbs free energy of unfolding (-0.5 ± 0.30) at 25°C and accordingly expected to be mostly unstructured *in vitro* at equilibrium [35].

In vitro, SPC-SH3 domains fold following a two-state process [33]. Assuming that the same mechanism applies for the folding of these domains in bacteria, partially folded intermediates would not accumulate significantly during the acquisition of the native structure after the biosynthesis at the ribosome. This implies that, if, as usually assumed, aggregation occurs during the folding reaction, one should expect the degree of protein deposition to correlate with the folding rate of the protein, since folding into the native structure through the establishment of specific intramolecular contacts and aggregation through anomalous intermolecular interactions are expected to compete kinetically [42].

In Fig. 2 we show the correlation between the observed experimental solubilities and the different calculated energy barriers to refolding as derived from stopped-flow experiments. It is important to note that the rate of protein folding (k_f) changes in an exponential manner with the refolding activation free energy. We observe a significant correlation between the energy barrier to refolding and solubility ($r = 0.69$ and $p < 0.002$). However, this correlation is clearly lower than that found for the thermodynamic stability, which suggests that it might be just a consequence of thermodynamic stabilities and folding rates being somehow linked in our protein set. Figure 2B demonstrates that this is likely the case, since, as a trend, more stable mutants tend to exhibit faster refolding rates as a result of the stabilization of

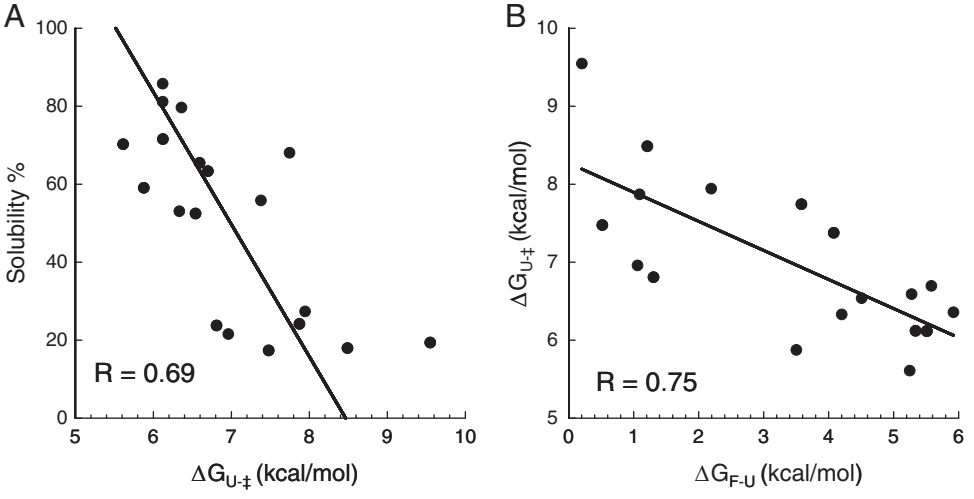


Figure 2. Relationship between folding activation free energy, thermodynamic stability and *in vivo* solubility. (A) Correlation between the energy barriers to folding and the *in vivo* solubility of SPC-SH3 domains. Solubility is expressed as the experimentally determined percentage of SH3 domain located in the soluble cell fraction relative to the total recombinant protein. Folding kinetic parameters were calculated at 25°C, pH 7.0, in 50 mM sodium phosphate. (B) Correlation between the folding activation-free energies and the conformational stability of SPC-SH3 domains.

the transition state ensemble and the consequent decrease of the energy barrier to folding, whereas an inverse tendency is observed for destabilized mutants. The behavior of C8A and its Ile to Val mutants C8A-I25V and C8A-I53V provide further support to the notion that folding rates do not control the aggregation of SH3 domains in the bacterial background. The Ile to Val mutations remove only a single methyl group but have important effects on the folding kinetics, decreasing the refolding rates relative to that of the original C8A protein [35]. Still, both mutants turn to be more soluble than C8A. Also, MAXF and Best7 exhibit faster refolding rates than the WT domain, while being clearly more aggregation-prone.

Protein aggregation usually follows a second-order reaction and is therefore very sensitive to protein concentration [43]. The data suggest that even at the high concentrations attained during the heterologous expression of SPC-SH3 domains, the aggregation reaction cannot kinetically compete with the fast attainment of the native structure that exhibit these small globular domains. During the revision of the present work Eichmann *et al.* used NMR spectroscopy to demonstrate that, in fact, the folding of recombinant SPC-SH3 occurs cotranslationally in bacteria, following a two-state transition without significant population of compact folding intermediates [30]. As proposed by the authors, it is likely that the ribosome may also protect emerging SPC-SH3 domains from aberrant interaction with cytosolic molecules. Therefore, SPC-SH3 molecules emerge from the ribosome as already folded and monomeric globular domains, in agreement with our observation that the energy barriers to refolding are not relevant for the intracellular aggregation of SH3 domains. Of course, this might not be the case for larger and more complex proteins whose folding pathways would be slower and involve the population of partially folded intermediates or the exposure of unstructured regions during their biosynthesis at the ribosome. To prevent the aggregation of such polypeptides the cell has evolved sophisticated molecular mechanisms including the ribosome-borne folding activity and the trigger factor, a chaperone interacting with nascent polypeptide chains as they protrude from the bacterial ribosome [44].

3.5 Correlation between energy barriers to unfolding and *in vivo* aggregation

Assuming that, in our protein data set, aggregation does not occur significantly during the attainment of the folded conformation in normal cell conditions, the fact that, at least *in* vitro at 25°C, most proteins have their equilibrium shifted towards the folded state, as indicated by their positive Gibbs free energy of unfolding, implies that the acquisition of a compact conformation does not completely protect SH3 domains from aggregation. In the cell, irreversible processes, like aggregation, might displace the folding–unfolding reaction from equilibrium, in such a way that a positive value for the unfolding Gibbs free energy might not necessarily guarantee that the protein will remain in the native state in a physiological relevant timescale [45]. As stated in the Introduction, for a number of proteins, kinetic stability instead of thermodynamic stability seems to control the aggregation process *in vitro* [23–27].

In this scenario both partially and totally unfolded conformers might trigger aggregation processes. To test if a totally unstructured SPC-SH3 domain is an aggregation-competent species when expressed in *E. coli*, we took advantage of the study of Eichmann *et al.*, which demonstrates by NMR in solution that a SPC-SH3 mutant lacking the five C-terminal residues displays a random-coil structure, either when bound to the ribosome or free in solution [30]. In Fig. 3 we show that this deletion mutant, although expressed at lower levels than the rest of variants, accumulates almost exclusively in the insoluble fraction, indicating that while partially folded intermediates might exist under certain conditions even for two-state folders [37], in the case of SPC-SH3 domains, their population is not obligatory for them to become insoluble.

For the present analysis we assume here that for two-state folding SH3 domains, aggregation in the cell occurs only from the unfolded state ensemble. Also, the aggregation of SH3 domains into IBs is considered to be irreversible, although the presence of chaperones might extract and assist the folding of the polypeptide chains embedded in the IBs [41]. Under these assumptions one might propose

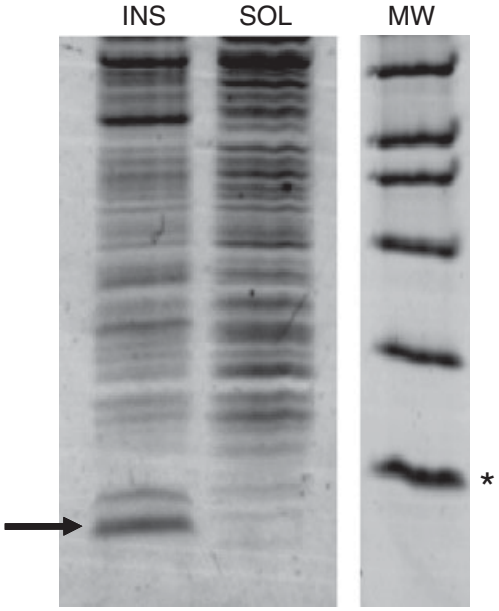


Figure 3. SDS-PAGE analysis of the intracellular solubility of an unstructured SPC-SH3 mutant. Soluble (SOL) and insoluble (INS) fractions of *E. coli* cells expressing a truncated SPC-SH3 domain lacking the five C-terminal residues [30]. The location of the SH3 domain is indicated by an arrow. The star indicates aprotinin (6500 Da).

the following Lumry–Eyring model [46] to describe the intracellular aggregation of SH3 domains:

$$\text{native state} \underset{k_f}{\overset{k_u}{\leftrightarrow}} \text{unfolded} \xrightarrow{k_a} \text{aggregated protein}$$

where k_u, k_f and k_a represent the unfolding, folding and aggregation rates, respectively.

The term k_u reflects the difference between the free energies of the native state and the transition state ensemble or, in more general terms, the kinetic stability of the protein. In this model the k_u and k_a values determine whether it is the thermodynamic or the kinetic stability that controls the level of soluble protein. In this scenario, the overall aggregation process cannot be faster than the unfolding rate (k_u), no matter how fast the aggregation rate is [45]. The maximum aggregation in a given timescale occurs when aggregation is fast enough to cause protein unfolding becoming the rate-limiting step of the overall aggregation process. In this condition an exact correlation between the levels of aggregated protein and kinetic stability is expected.

Figure 4 shows that there is a high and significant correlation ($r = 0.84$ and $p < 0.00001$) between the observed experimental solubilities and the different kinetic stability for the SH3 domains in this study as calculated from stopped-flow experiments. Again, this correlation is lower than that found for thermodynamic stability, and the relationship between kinetic stability and aggregation might result from the association between thermodynamic and energy barriers to unfolding, as shown in Fig. 4B, rather than reflecting the effect of unfolding rates on IB formation. Analysis of the unfolding properties of some of the mutants allows deciding between these two possibilities. C8A unfolds ≈30-fold faster than the WT protein while displaying very similar solubility inside bacteria. D48G unfolds ≈3 times faster than WT but displays higher solubility. The unfolding rates of Best9, Best7 and MaxF are ≈20 time higher than that of MAXL and Best2 while displaying similarly low solubilities.

Overall, the results strongly support the thermodynamic stability as the key factor controlling the intracellular aggregation of SH3 domains in our model system. According to our simplified scheme, this implies that the aggregation rate is exceedingly slow relative to the unfolding rates of the mutants. Under these conditions aggregation becomes almost independent of the individual unfolding and refolding rates and it is the concentration of the unfolded, assembly-prone species at equilibrium or, in other words, the conformational stability, which modulates the level of aggregation. Importantly, although unspecific aggregation is known to occur rapidly, accumulating evidences indicate that protein aggregation in bacteria relies on the establishment of sequence-specific intermolecular interactions leading to the formation of intermolecular β-sheets displaying amyloid-like structures [9], a process that might exhibit significant lag phases until the formation of effective nuclei occurs and polymerization can proceed with a high rate, as demonstrated for many amyloid fibrils *in vitro*.

3.6 Correlation between thermal stability and the aggregation of SH3 domains in cytosolic cell extracts

A widespread strategy to reduce the aggregation of recombinant polypeptides relies on cell growth at reduced temperatures [47]. Accordingly, when aggregation-prone

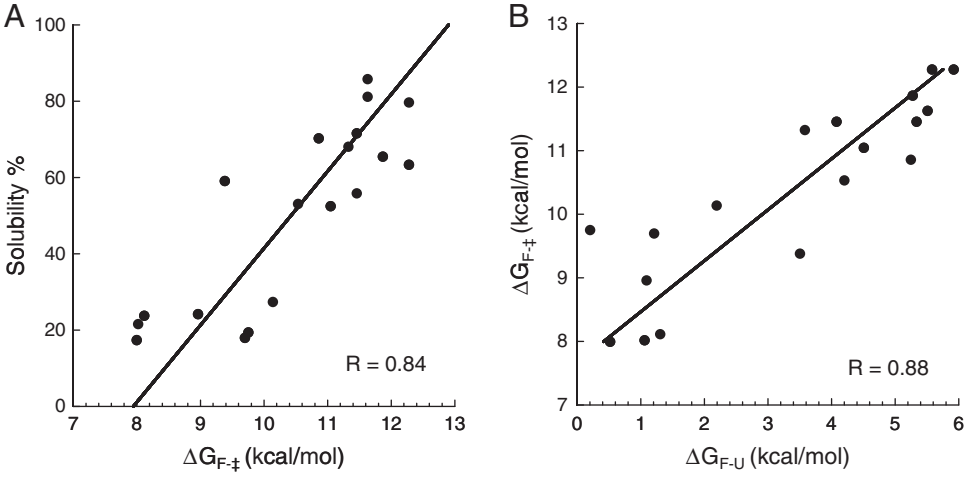


Figure 4. Relationship between kinetic stability, thermodynamic stability and *in vivo* solubility. (A) Correlation between the kinetic stability and the *in vivo* solubility of SPC-SH3 domains. Solubility is expressed as the experimentally determined percentage of SH3 domain located in the soluble cell fraction relative to the total recombinant protein. Folding kinetic parameters were calculated at 25°C, pH 7.0, in 50 mM sodium phosphate. (B) Correlation between the kinetic stability and the conformational stability of SPC-SH3 domains.

mutants like MAXF or MAXI are expressed at 18°C the amount of protein in the soluble fraction increases from ≈25 to ≈50%. It has been assumed that this increase in solubility results mainly from lower translation rates at the ribosome at low temperatures, which decreases the concentration of unfolded or partially folded species and shifts the competition between folding and aggregation towards the formation of native intramolecular interactions [47]. However, because for SPC-SH3 domains folding appears to occur cotranslationally [30] and folding rates accordingly appear to have a minor influence on the level of cellular aggregated protein, even at the high translation rates attained at 37°C, a decrease in translation rates is expected to contribute little to the higher level of solubility of these domains when expressed at low temperature.

The effect of temperature on the solubility of SPC-SH3 domains can be rationalized if we consider that the final level of recombinant aggregated protein inside bacteria depends mainly on the unfolded domain population at equilibrium and on the aggregation rate from this state, since temperature affects both parameters. Aggregation rates are known to be highly dependent on the temperature and, as a general rule, a decrease in temperature results in reduced aggregation rates, as observed for many amyloidogenic proteins [43, 48]. In addition, in general, folded proteins become more stable at low temperature because the entropy term TΔS which favors the unfolded state, becomes smaller resulting in an increase in the Gibbs free energy to unfolding ($\Delta G_u = \Delta H_u - T\Delta S_u$) and a higher ratio of folded to unfolded conformations at equilibrium. Both effects would contribute to increase the solubility of SH3 domains in a given timescale and environment.

To study the effect of temperature on the aggregation of SH3 proteins, four domains exhibiting different solubility and stability, namely MAXF, MAXI, WT and C8A-I25V were expressed in *E. Coli* at 18°C. After a ten-fold concentration of the culture, the cells were lysated and the complete soluble cell fraction containing the different SH3 domains together with all the cytosolic protein content was recovered. The concentration of SH3 domains in this cell lysate is ≈5 mg/mL. The different cell lysates were incubated at 18, 37 or 50°C for 16 h and the amount of SH3 domains remaining soluble visualized on Tricine-SDS gels (Fig. 5). All the mutants kept soluble when incubated at 18°C. At 37°C the destabilized mutants MAXF and MAXI aggregated significantly, with ≈75% of the protein becoming insoluble in this condition. These two mutants become almost totally aggregated when incubated at 50°C. In contrast, the WT protein and the overstabilized C8A-I25V mutant remained essentially soluble even when incubated at 50°C. The thermal unfolding of these mutants was analyzed by monitoring changes in the distinctive SPC-SH3 circular dichroism band at 220 nm [20]. All mutants unfolded in a cooperative manner, allowing the calculation of the associated melting Tm of 33±0.9, 42±0.6, 65±0.2 and 71±0.1°C for MAXF, MAXI, WT and C8A-I25V, respectively. This differential thermal stabilities are in good agreement with their relative stabilities in front of chemical denaturation [35]. According to their Tm, both MAXF and MAXI would present a significant population of unfolded conformers at 37°C, which would explain why they already aggregate in the cell lysate at this physiological temperature. The high thermal stability of WT and C8A-I25V would prevent their aggregation at 50°C.

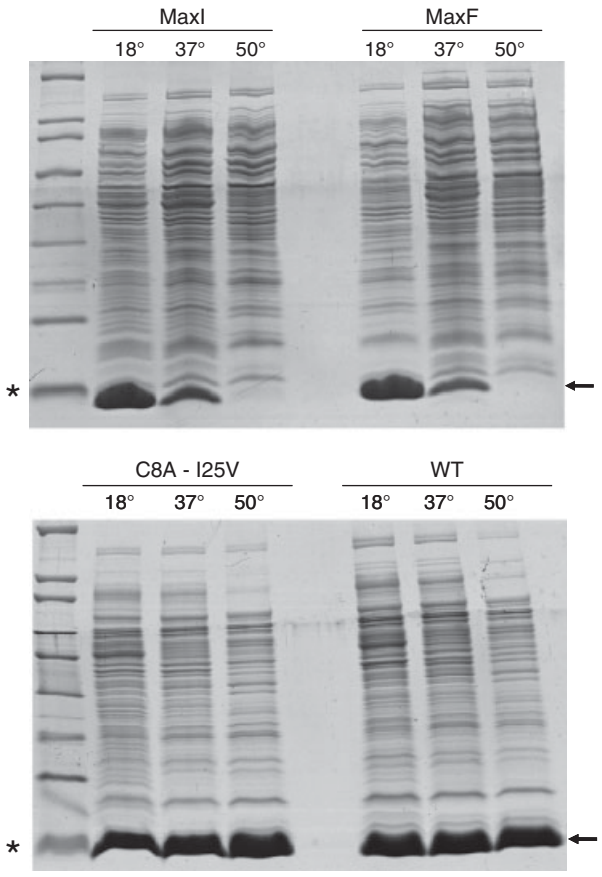


Figure 5. Temperature-induced aggregation of SH3 domains in cell lysates. The cytosolic fractions of cells expressing the different domains at 18°C were incubated for 16 h at the indicated temperature, centrifuged and the soluble fraction resolved on Tricine- SDS/12% (w/v) PAGE gels. The location of SH3 domains is indicated by an arrow. The star indicates aprotinin (6500 Da).

Shifting the temperature to non-physiological values changes the protein's energy landscape. Therefore, we decided to follow the kinetics of aggregation of the SH3 domains in the cell lysate at 37°C, the same temperature used for bacterial culture and protein production (Fig. 6). The dependence of the fraction of aggregated protein on the reaction time could be fitted to a single constant first-order reaction ($R > 0.96$) for MAXF and MAXI, with rate constants of 6.7×10^{-5} and $1.7 \times 10^{-5}\,s^{-1}$, respectively, in good agreement with their relative thermal stabilities.

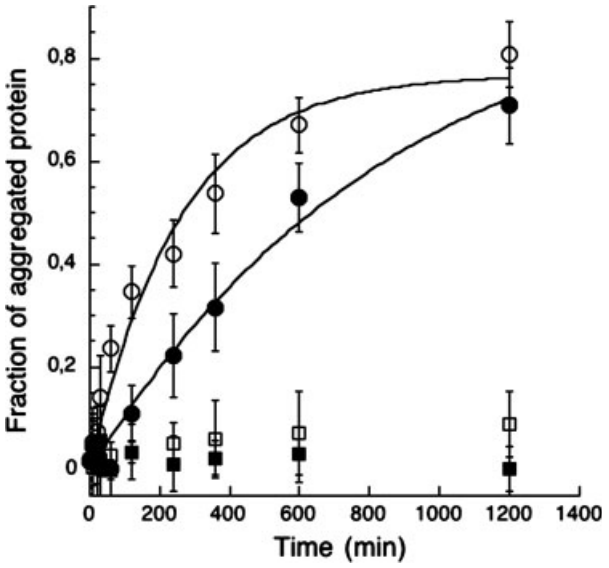


Figure 6. Aggregation kinetics of SPC-SH3 domains in cell lysates at 37°C. The cytosolic fractions of cells expressing MAXF (empty circles), MAXI (solid circles), WT (empty squares) and C8A-I25V (solid squares) domains at 18°C were incubated at 37°C and the percentage of SH3 domain remaining soluble quantified at each time point on Tricine-SDS/12% (w/v) PAGE gels. The aggregation reactions of MAXF and MAXI were fitted to first-order reactions to calculate the corresponding rate constants. Standard errors for aggregation values are shown.

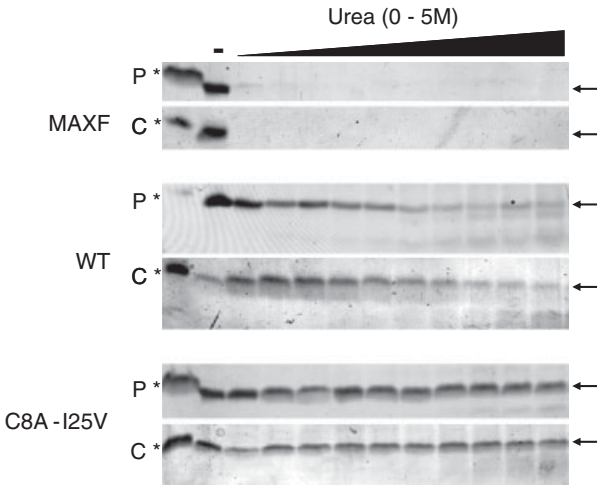


Figure 7. *In vitro* and in *situ* analysis of SPC-SH3 domains conformational stability by pulse proteolysis. Tricine-SDS/12% (w/v) PAGE gels of (P) purified SH3 domains and (C) cytosolic fractions of cells expressing SH3 domains equilibrated overnight at the indicated urea concentrations and pulse treated with thermolysin. An arrow indicates the location of SH3 domains. The star indicates aprotinin (6500 Da). In the negative control line (–) the protein was incubated in the absence of urea and thermolysin.

Although proteins unfolding in response to chemically and thermally induced perturbations may be exposed to very different unfolding energy landscapes, in the case of SPC-SH3 domains it has been shown *in vitro* that, under both conditions, these domains follow a highly cooperative two-state unfolding mechanism, without evident accumulation of partially folded intermediates. Therefore, although the presence of other species in cell lysates cannot be completely discarded, the population of unfolded, aggregation-prone, conformers at equilibrium can be postulated as the most probable precursor of the insoluble species.

WT and C8A-I25V display lower aggregation propensity in the cell lysate than in intact bacteria, where a fraction of these domains is found as IBs. Although cell lysates mimic better than a simple buffer the cytosolic environment, the significant dilution of the cellular components, including the recombinant domains, would result in lower aggregation rates than those expected to occur in intact cells.

3.7 Correlation between the stability against chemical denaturation of SH3 domains *in vitro* and in cell lysates

A common assumption in biochemistry is that the conformational stability of proteins in simple *in vitro* conditions is similar to that displayed in the complex intracellular environment. To provide evidence for this assumption we compared the stability of different SH3 domains submitted to chemical denaturation *in vitro* and in bacterial cytosolic extracts using pulse proteolysis. Native-state proteolysis allows investigating the thermodynamic accessibilities of partially or fully unfolded forms (cleavable forms) under equilibrium conditions both *in vitro* and in cell lysates [32, 49]. Purified MAXF, WT and C8A-I25V domains as well as the lysates of cells expressing these variants were equilibrated in the presence of increasing concentrations of urea at 18°C, digested with thermolysin at 37°C for 2 min and the amount of protease resistant SH3 domain at the different chaotropic agent concentrations was analyzed on Tricine-SDS gels (Fig. 7). Unfortunately, the presence of urea in this type of gels promoted broadening and diffusion of the SH3 bands, precluding accurate quantification and therefore calculation of the precise Cm for proteolysis. Nevertheless, we could confirm that the qualitative resistance of SH3 domains to proteolysis *in vitro* and cell lysates are similar, as previously described for the small GTPase protein H-ras [49]. The unstable mutant MAXF is almost totally digested even in the absence of urea, supporting the notion that a significant population of this domain is unfolded at physiological temperature, and therefore susceptible of aggregation as shown in the previous section. The over-stabilized C8A-I25V mutant essentially remains resistant to proteolysis under all the conditions of the assay whereas the WT displays an intermediate stability in front of proteolysis and remains resistant at moderate urea concentrations. Therefore, as a trend, the stabilities of SH3 against chemical denaturation *in vitro* and in the more complex cell lysate appear to be associated.

As discussed above, the apparent inertness of cytosolic proteins on the stability of SH3 domains does not necessarily apply *in vivo* since the cellular context differs significantly from that in the dilute cell lysate. The much more crowded environment inside the cell may condition the conformational energetics of proteins [50]. Nevertheless, in principle, one should not expect the effect of macromolecular crowding being more important for some of the proteins in our set than for others and therefore the relative stabilities of the different proteins *in vitro* or in the cytosolic extracts are expected to resemble, at least in relative terms and for this particular SH3 domain set, those in the cellular background.

3.8 Correlation between the stability of SH3 domains against proteolysis under non-denaturing conditions *in vitro* and in cell lysates

Chemical denaturation is perhaps the most widely used technique to approximate protein thermodynamic stability. However, the denaturation of a protein by chaotropic agents may not necessarily reflect the unfolding of the polypeptide in the native cell conditions, even when complete cell lysates are used, like in the present work. To study protein stability under near to native conditions we followed the kinetics of proteolysis by thermolysin at 37°C in the absence of denaturants. We analyzed the proteolytic stability of purified MAXF, WT and C8A-I25V proteins as well as of cell lysates containing these domains (Fig. 8). Again, a qualitative agreement between the *in vitro* resistance to proteolitic digestion and that in cell lysates is observed for the three SPC-SH3 domains. The destabilized mutant MAXF is almost totally digested after 1 min of reaction, whereas WT and specially the over-stabilized C8A-I25V exhibit a much lower digestion rate in both purified and crude samples. Therefore, a qualitative association between the resistance of a particular SH3 domain in front of chemical denaturation and its resistance to proteolysis under native-like conditions is observed, suggesting that the thermodynamic stability of the domain is a critical factor controlling the exposure of natively protected thermolysin cleavable sequences.

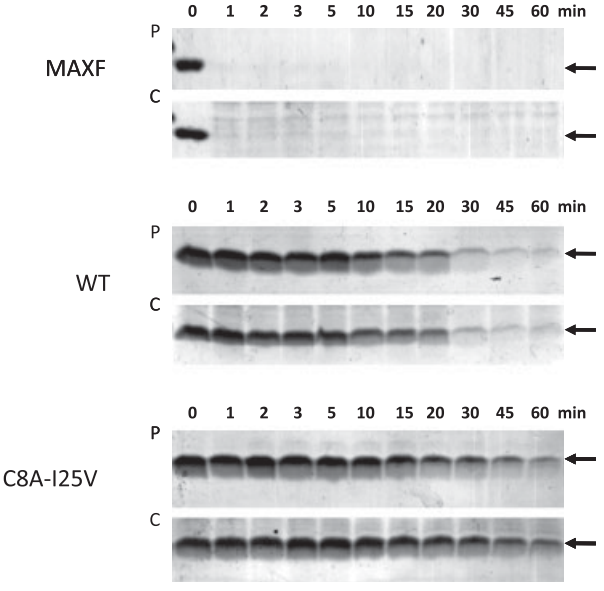


Figure 8. *In vitro* and in *situ* analysis of SPC-SH3 domains proteolytic resistance under native-like conditions. Tricine-SDS/12% (w/v) PAGE gels of time-course proteolysis experiments under non-denaturing conditions. (P) purified SH3 domains and (C) cytosolic fractions of cells expressing SH3 domains were incubated at 37°C in the presence of thermolysin for the indicated time.

4 Concluding remarks

The increasing evidence for the formation of amyloid-like structures in *E. coli* makes this organism a tractable system to understand the molecular determinants shaping protein aggregation in the cell. Here we have exploited this system to discriminate whether the deposition of small globular proteins in the bacterial cytosol is dependent on the thermodynamic or the kinetic stability, using the well-characterized SPC-SH3 domain as a model system. This constitutes one of the few studies experimentally addressing how *in vitro* folding and unfolding rates correlate with intracellular protein aggregation. Overall, the present results allow us to propose a model for the aggregation of recombinant SH3 domains in the bacterial cytosol. The connection between conformational stability and aggregation in SH3 domains is likely a consequence of the folded and unfolded states being in rapid pre-equilibrium relative to the rate of aggregation. Under these conditions, self-assembly reactions become mainly sensitive to the population of unfolded molecules in equilibrium with the native state for each particular sequential variant. The *in vitro* folding and unfolding rates of many small globular proteins are known or expected to be in the same range as that of SPC-SH3 [51]. In addition, the cellular concentrations in normal physiological conditions are likely much lower than those in the present study for most of these proteins. Taken together these evidences suggest that the thermodynamic stability might be an important factor modulating the intracellular aggregation of small domains displaying a fast folding process that follows a two-state mechanism.

In principle, if solubility and stability are associated, a high conformational stability would constitute a selective advantage since the anomalous self-assembly of polypeptides has, in most cases, deleterious effects for the cell. Then, the question arises: Why do natural proteins display only a marginal stability of few kcal/mol? On the one hand, there is growing evidence that protein stabilization can be detrimental to function [52, 53]. On the other hand, if aggregation rates are small enough, even marginal stabilities might provide protection against protein aggregation in a physiological relevant timescale. Recent works by Tartaglia and Vendruscolo [54–56] as well as by our group

[57] indicate that highly abundant proteins display sequential features that decrease their aggregation propensity from unfolded states. This evolutionarily developed protection mechanism suggests that the concentration of abundant proteins is probably high enough for their aggregation rates to be relevant in the cellular environment and timescale, in such a way that the presence of high-energy conformations, including partially or globally unfolded species, in dynamic equilibrium with the native conformation might become a danger. To the best of our knowledge, the specific stability features of high- and low-abundant proteins have not been addressed yet. In principle, one might also expect that highly abundant proteins display higher stabilities to prevent their aggregation, but it is likely that the deleterious effects of high stability and conformational rigidity on protein activity would largely counterbalance this effect. In such cases, a high kinetic stability might provide protection against aggregation without a need for more stable but inactive conformations [58].

There is an increasing interest in developing efficient algorithms to predict the aggregation of proteins in living organisms. To a first approach, without taking into account the important role of the cellular quality control machinery on the protein folding homeostasis, the aggregation of small globular domains in the cell might be modeled as a combination of the intrinsic aggregation propensity of the polypeptide unfolded state as determined by the sequence, the population of the unfolded state at equilibrium as determined by the *in vivo* thermodynamic stability and the aggregation rate in this environment as determined by the protein concentration and solution conditions. Therefore, it is suggested that the introduction of stability features should be seriously considered in algorithms aimed to predict the aggregation propensities of globular proteins in their physiological or heterologous environments. Importantly, the firsts steps towards this direction have been already taken and shown to increase the accuracy of aggregation predictions for structured proteins [59].

This work was supported by grants BIO2007-68046 from Ministerio de Ciencia e Innovación (Spain) and 2009-SGR 760 from AGAUR (Generalitat de Catalunya). S.V. has been granted an ICREA ACADEMIA award (ICREA).

The authors have declared no conflict of interest.

5 References

[1] Chiti, F., Dobson, C. M., Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.* 2006, *75*, 333–366.

[2] Fernandez-Busquets, X., de Groot, N. S., Fernandez, D., Ventura, S., Recent structural and computational insights into conformational diseases. *Curr. Med. Chem.* 2008, *15*, 1336–1349.

[3] Soto, C., Estrada, L. D., Protein misfolding and neurodegeneration. *Arch. Neurol.* 2008, *65*, 184–189.

[4] Pepys, M. B., Amyloidosis. *Annu. Rev. Med.* 2006, *57*, 223–241.

[5] Pallares, I., Vendrell, J., Aviles, F. X., Ventura, S., Amyloid fibril formation by a partially structured intermediate state of alpha-chymotrypsin. *J. Mol. Biol.* 2004, *342*, 321–331.

[6] Guijarro, J. I., Sunde, M., Jones, J. A., Campbell, I. D., Dobson, C. M., Amyloid fibril formation by an SH3 domain. *Proc. Natl. Acad. Sci. USA* 1998, *95*, 4224–4228.

[7] Ventura, S., Zurdo, J., Narayanan, S., Parreno, M. *et al.*, Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case. *Proc. Natl. Acad. Sci. USA* 2004, *101*, 7258–7263.

[8] Chiti, F., Taddei, N., Bucciantini, M., White, P. *et al.*, Mutational analysis of the propensity for amyloid formation by a globular protein. *EMBO J.* 2000, *19*, 1441–1449.

[9] Morell, M., Bravo, R., Espargaro, A., Sisquella, X. *et al.*, Inclusion bodies: Specificity in their aggregation process and amyloid-like structure. *Biochim. Biophys. Acta* 2008, *1783*, 1815–1825.

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

[11] de Groot, N. S., Sabate, R., Ventura, S., Amyloids in bacterial inclusion bodies. *Trends Biochem. Sci.* 2009, *34*, 408–416.

[12] Wasmer, C., Benkemoun, L., Sabate, R., Steinmetz, M. O. *et al.*, Solid-state NMR spectroscopy reveals that E. coli inclusion bodies of HET-s(218-289) are amyloids. *Angew. Chem. Int. Ed. Engl.* 2009, *48*, 4858–4860.

[13] Wang, L., Maji, S. K., Sawaya, M. R., Eisenberg, D., Riek, R., Bacterial inclusion bodies contain amyloid-like structure. *PLoS Biol.* 2008, *6*, e195.

[14] Uversky, V. N., Fink, A. L., Conformational constraints for amyloid fibrillation: the importance of being unfolded. *Biochim Biophys Acta* 2004, *1698*, 131–153.

[15] Castillo, V., Ventura, S., Amyloidogenic regions and interaction surfaces overlap in globular proteins related to conformational diseases. *PLoS Comput. Biol.* 2009, *5*, e1000476.

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

[17] Winkelmann, J., Calloni, G., Campioni, S., Mannini, B. *et al.*, Low-level expression of a folding-incompetent protein in *Escherichia coli*: search for the molecular determinants of protein aggregation *in vivo*. *J. Mol. Biol.* 2010, *398*, 600–613.

[18] Calloni, G., Zoffoli, S., Stefani, M., Dobson, C. M., Chiti, F., Investigating the effects of mutations on protein aggregation in the cell. *J. Biol. Chem.* 2005, *280*, 10607–10613.

[19] Mayer, S., Rudiger, S., Ang, H. C., Joerger, A. C., Fersht, A. R., Correlation of levels of folded recombinant p53 in *Escherichia coli* with thermodynamic stability in vitro. *J. Mol. Biol.* 2007, *372*, 268–276.

[20] Espargaro, A., Castillo, V., de Groot, N. S., Ventura, S., The *in vivo* and *in vitro* aggregation properties of globular proteins correlate with their conformational stability: the SH3 case. *J. Mol. Biol.* 2008, *378*, 1116–1131.

[21] Godoy-Ruiz, R., Ariza, F., Rodriguez-Larrea, D., Perez-Jimenez, R. *et al.*, Natural selection for kinetic stability is a likely origin of correlations between mutational effects on protein energetics and frequencies of amino acid occurrences in sequence alignments. *J. Mol. Biol.* 2006, *362*, 966–978.

[22] Fersht, A. R., Sato, S., Phi-value analysis and the nature of protein-folding transition states. *Proc. Natl. Acad. Sci. USA* 2004, *101*, 7976–7981.

[23] Foss, T. R., Wiseman, R. L., Kelly, J. W., The pathway by which the tetrameric protein transthyretin dissociates. *Biochemistry* 2005, *44*, 15525–15533.

[24] Guo, Z., Eisenberg, D., The mechanism of the amyloidogenic conversion of T7 endonuclease I. *J. Biol. Chem.* 2007, *282*, 14968–14974.

[25] Ellisdon, A. M., Thomas, B., Bottomley, S. P., The two-stage pathway of ataxin-3 fibrillogenesis involves a polyglutamine-independent step. *J. Biol. Chem.* 2006, *281*, 16888–16896.

[26] Liemann, S., Glockshuber, R., Influence of amino acid substitutions related to inherited human prion diseases on the thermodynamic stability of the cellular prion protein. *Biochemistry* 1999, *38*, 3258–3267.

[27] Knaupp, A. S., Levina, V., Robertson, A. L., Pearce, M. C., Bottomley, S. P., Kinetic instability of the serpin Z alpha1-antitrypsin promotes aggregation. *J. Mol. Biol.* 396, 375–383.

[28] Martinez, J. C., Serrano, L., The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nat. Struct. Biol.* 1999, *6*, 1010–1016.

[29] Ventura, S., Vega, M. C., Lacroix, E., Angrand, I. *et al.*, Conformational strain in the hydrophobic core and its implications for protein folding and design. *Nat. Struct. Biol.* 2002, *9*, 485–493.

[30] Eichmann, C., Preissler, S., Riek, R., Deuerling, E., Cotranslational structure acquisition of nascent polypeptides monitored by NMR spectroscopy. *Proc. Natl. Acad. Sci. USA* 107, 9111–9116.

[31] Martinez, J. C., Pisabarro, M. T., Serrano, L., Obligatory steps in protein folding and the conformational diversity of the transition state. *Nat. Struct. Biol.* 1998, *5*, 721–729.

[32] Park, C., Marqusee, S., Pulse proteolysis: a simple method for quantitative determination of protein stability and ligand binding. *Nat. Methods* 2005, *2*, 207–212.

[33] Viguera, A. R., Martinez, J. C., Filimonov, V. V., Mateo, P. L., Serrano, L., Thermodynamic and kinetic analysis of the SH3 domain of spectrin shows a two-state folding transition. *Biochemistry* 1994, *33*, 2142–2150.

[34] Martinez, J. C., Serrano, L., The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nat. Struct. Biol.* 1999, *6*, 1010–1016.

[35] Ventura, S., Vega, M. C., Lacroix, E., Angrand, I. *et al.*, Conformational strain in the hydrophobic core and its implications for protein folding and design. *Nat. Struct. Biol.* 2002, *9*, 485–493.

[36] Dobson, C. M., Principles of protein folding, misfolding and aggregation. *Semin. Cell Dev. Biol.* 2004, *15*, 3–16.

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

[38] Booth, D. R., Sunde, M., Bellotti, V., Robinson, C. V. *et al.*, Instability, unfolding and aggregation of human lysozyme variants underlying amyloid fibrillogenesis. *Nature* 1997, *385*, 787–793.

[39] Smith, D. P., Jones, S., Serpell, L. C., Sunde, M., Radford, S. E., A systematic investigation into the effect of protein destabilisation on beta 2-microglobulin amyloid formation. *J. Mol. Biol.* 2003, *330*, 943–954.

[40] de Groot, N. S., Espargaro, A., Morell, M., Ventura, S., Studies on bacterial inclusion bodies. *Future Microbiol.* 2008, *3*, 423–435.

[41] Ventura, S., Villaverde, A., Protein quality in bacterial inclusion bodies. *Trends Biotechnol.* 2006, *24*, 179–185.

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

[43] Morris, A. M., Watzky, M. A., Finke, R. G., Protein aggregation kinetics, mechanism, and curve-fitting: a review of the literature. *Biochim. Biophys. Acta* 2009, *1794*, 375–397.

[44] Sabate, R., de Groot, N. S., Ventura, S., Protein folding and aggregation in bacteria. *Cell. Mol. Life Sci.* 2010, *67*, 2695–2715.

[45] Plaza del Pino, I. M., Ibarra-Molero, B., Sanchez-Ruiz, J. M., Lower kinetic limit to protein thermal stability: a proposal regarding protein stability *in vivo* and its relation with misfolding diseases. *Proteins* 2000, *40*, 58–70.

[46] Sanchez-Ruiz, J. M., Theoretical analysis of Lumry-Eyring models in differential scanning calorimetry. *Biophys. J.* 1992, *61*, 921–935.

[47] Sorensen, H. P., Mortensen, K. K., Soluble expression of recombinant proteins in the cytoplasm of *Escherichia coli*. *Microb. Cell Fact.* 2005, *4*, 1.

[48] Sabate, R., Castillo, V., Espargaro, A., Saupe, S. J., Ventura, S., Energy barriers for HET-s prion forming domain amyloid formation. *FEBS J.* 2009, *276*, 5053–5064.

[49] Kim, M. S., Song, J., Park, C., Determining protein stability in cell lysates by pulse proteolysis and Western blotting. *Protein Sci.* 2009, *18*, 1051–1059.

[50] Minton, A. P., Implications of macromolecular crowding for protein assembly. *Curr. Opin. Struct. Biol.* 2000, *10*, 34–39.

[51] Maxwell, K. L., Wildes, D., Zarrine-Afsar, A., De Los Rios, M. A. *et al.*, Protein folding: defining a "standard" set of experimental conditions and a preliminary kinetic data set of two-state proteins. *Protein Sci.* 2005, *14*, 602–616.

[52] Varley, P. G., Pain, R. H., Relation between stability, dynamics and enzyme activity in 3-phosphoglycerate

kinases from yeast and *Thermus thermophilus*. *J. Mol. Biol.* 1991, *220*, 531–538.

[53] Shoichet, B. K., Baase, W. A., Kuroki, R., Matthews, B. W., A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci. USA* 1995, *92*, 452–456.

[54] Tartaglia, G. G., Pechmann, S., Dobson, C. M., Vendruscolo, M., Life on the edge: a link between gene expression levels and aggregation rates of human proteins. *Trends Biochem. Sci.* 2007, *32*, 204–206.

[55] Tartaglia, G. G., Pechmann, S., Dobson, C. M., Vendruscolo, M., A relationship between mRNA expression levels and protein solubility in *E. coli*. *J. Mol. Biol.* 2009, *388*, 381–389.

[56] Tartaglia, G. G., Vendruscolo, M., Correlation between mRNA expression levels and protein aggregation propensities in subcellular localisations. *Mol. Biosyst.* 2009, *5*, 1873–1876.

[57] de Groot, N. S., Ventura, S., Protein aggregation profile of the bacterial cytosol. *PLoS One* 2010, *5*, e9383.

[58] Cabrita, L. D., Bottomley, S. P., How do proteins avoid becoming too stable? Biophysical studies into metastable proteins. *Eur. Biophys. J.* 2004, *33*, 83–88.

[59] Tartaglia, G. G., Pawar, A. P., Campioni, S., Dobson, C. M. *et al.*, Prediction of aggregation-prone regions in structured proteins. *J Mol. Biol.* 2008, *380*, 425–436.

Amyloidogenic Regions and Interaction Surfaces Overlap in Globular Proteins Related to Conformational Diseases

Virginia Castillo, Salvador Ventura*

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

Abstract

Protein aggregation underlies a wide range of human disorders. The polypeptides involved in these pathologies might be intrinsically unstructured or display a defined 3D-structure. Little is known about how globular proteins aggregate into toxic assemblies under physiological conditions, where they display an initially folded conformation. Protein aggregation is, however, always initiated by the establishment of anomalous protein-protein interactions. Therefore, in the present work, we have explored the extent to which protein interaction surfaces and aggregation-prone regions overlap in globular proteins associated with conformational diseases. Computational analysis of the native complexes formed by these proteins shows that aggregation-prone regions do frequently overlap with protein interfaces. The spatial coincidence of interaction sites and aggregating regions suggests that the formation of functional complexes and the aggregation of their individual subunits might compete in the cell. Accordingly, single mutations affecting complex interface or stability usually result in the formation of toxic aggregates. It is suggested that the stabilization of existing interfaces in multimeric proteins or the formation of new complexes in monomeric polypeptides might become effective strategies to prevent disease-linked aggregation of globular proteins.

Citation: Castillo V, Ventura S (2009) Amyloidogenic Regions and Interaction Surfaces Overlap in Globular Proteins Related to Conformational Diseases. PLoS Comput Biol 5(8): e1000476. doi:10.1371/journal.pcbi.1000476

Editor: Ruth Nussinov, National Cancer Institute, United States of America and Tel Aviv University, Israel

Received March 4, 2009; **Accepted** July 20, 2009; **Published** August 21, 2009

Funding: This work has been supported by grant BIO2007-68046 (Ministerio de Ciencia e Innovacion, Spain). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: salvador.ventura@uab.es

Introduction

The formation of insoluble amyloid protein deposits in tissues is related to the development of more than 40 different human diseases, many of which are debilitating and often fatal. The polypeptides responsible for these disorders are not related in terms of sequence or conformation [1–6]. Some of these proteins and peptides are mostly unstructured. Examples include amylin, amyloid-β-protein and α-synuclein. In contrast, many other amyloidogenic proteins are globular in their native state, implying that they have a properly packed and cooperatively sustained structure under physiological conditions. This group includes ß-2-microglobulin, transthyretin, lysozyme, superoxide dismutase 1 and immunoglobulins. As a general trend, evolution has endorsed globular proteins with solubility in their biological environments [7]. However, it has been shown that, *in vitro*, under conditions where they become totally or partially unfolded, both these pathogenic proteins [8–11] and many globular polypeptides not related to disease [12–15] readily convert into aggregates and ultimately into highly structured amyloid fibrils. This self-assembly process is triggered by the destabilization and opening of the native structure, which exposes previously protected aggregation-prone regions that can nucleate the aggregation reaction and participate in forming the β-core of the mature fibril through specific intermolecular interactions [16–18]. Such amyloidogenic sequence stretches have been described in most of the polypeptides underlying neurodegenerative and systemic amyloidogenic disorders. The main intrinsic protein properties that promote the assembly of such sequences into fibrils have been recently defined [19], and several algorithms that predict amyloidogenic sequences with good accuracy are already available [3,20,21].

Although the study of protein aggregation from non-native states has provided a wealth of data on the physico-chemical determinants of amyloid formation, little is known about how globular proteins aggregate from their initially folded and soluble conformations under physiological conditions, where extensive unfolding is not expected to occur [22]. Deciphering this issue is important because the deposition of globular polypeptides is linked to devastating disorders, and there is an urgent need for therapeutic intervention.

Protein aggregation can be seen as an anomalous type of protein-protein interaction. In functional interactions, binding partners come together in a stable and precise orientation in seconds [23]. This efficiency relies on the structural features of the interacting surfaces. Perhaps the most significant characteristic of a functional protein-protein interface is the presence of small high-affinity regions within the interface, with a reduced number of residues accounting for most of the binding energy [24–26]. Several computational approaches have been shown to forecast such regions with high accuracy [27–34]. Statistical analysis of the structures of protein-protein interfaces has revealed that tryptophan, phenylalanine, and methionine and to a lesser extent leucine, valine, and tyrosine are preferentially conserved at interaction sites [35]. The same residues have been shown to be conserved in the aggregation-prone sequences of the human proteome [36]. This suggests an intriguing possibility: that amyloidogenic regions and interacting surfaces might overlap in globular proteins. Several of the folded proteins linked to amyloid

Author Summary

The aggregation of proteins in tissues is associated with the pathogenesis of more than 40 human diseases. The polypeptides underlying disorders such as Alzheimer's and Parkinson's are devoid of any regular structure, whereas the polypeptides causing familial amyotrophic lateral sclerosis or nonneuropathic systemic amyloidosis correspond to globular proteins. Little is known about the mechanism by which globular proteins under physiological conditions aggregate from their initially folded and soluble conformations. Interestingly, several of these pathogenic proteins display quaternary structure or are bound to other proteins in their physiological context. In the present work, we show that protein-protein interaction surfaces and regions with high aggregation propensity significantly overlap in these polypeptides. This suggests that the formation of native complexes and self-aggregation reactions probably compete in the cell, explaining why point mutations affecting the interface or the stability of the protein complex lead in many cases to the formation of toxic aggregates. This study proposes general strategies to fight against diseases associated with the deposition of globular polypeptides.

diseases display quaternary structure or are bound to other proteins in their physiological context. If these interactions specifically cover amyloidogenic regions, they could play a role in protecting native-state proteins from aggregation. Alternatively, incorrect docking of interfaces might facilitate the assembly of overlapping amyloidogenic regions and therefore the formation of toxic protein aggregates of globular proteins. In the present work, we have used available computational approaches to predict aggregation-prone sequences and interacting residues in order to assess the extent to which these regions coincide in pathogenic and non-pathogenic proteins.

Results/Discussion

Prediction of Aggregation-Prone Regions and Protein-Protein Interaction Sites

The prediction of regions responsible for aggregation based on the primary sequence of a protein has been tackled by several methods, from simple considerations of the properties of amino acids to complex molecular dynamics calculations [37–44]. Overall, most of these methods predict with reasonable precision the regions of proteins in the cross-ß core of amyloid fibrils. This accuracy allows the proposal that the aggregation propensity of a polypeptide chain is ultimately dictated by the sequence [45]. Here we have used four different algorithms in parallel to provide a consensus prediction of the amyloidogenic regions in globular proteins linked to deposition diseases (see Methods). We chose the algorithms implemented by Fernandez-Escamilla *et al.* (TANGO) [38], Conchillo-Sole *et al.* (AGGRESCAN) [40], Galzitskaya *et al.* [41], and Zhang *et al.* [43]. All of them use the primary sequence as input and assume that the detected regions need to be at least partially exposed to solvent in order to nucleate the aggregation reaction.

Identification of binding sites in polypeptides is a direct computational approach to deciphering biological and biochemical function. Although sequence-based approaches to identifying protein interfaces exist, their results are often unsatisfactory. Here, we have used three different structure-based methods whose algorithms are publicly available as web servers to produce a consensus prediction of the interaction interfaces in the globular proteins under consideration (see Methods). These structure-based methods were developed by Fernandez-Recio *et al.* (ODA) [32], Murakami and Jones (SHARP^2) [31], and Negi *et al.* (InterProSurf) [33]. Although they are based on different principles and implement diverse computational strategies, all of them use the unbound three-dimensional structure of a globular protein as input.

Two levels of prediction were considered: i) residues predicted or shown to be both in aggregation-prone regions and at interfaces and ii) residues in aggregation-prone sequences that are close in space to the interaction surface (below 3 Å). The interaction predictions were compared with the experimentally determined contacts in the quaternary structure of the proteins or in complexes of the studied proteins with other polypeptides. The regions predicted to have high aggregation propensity were compared with fragments of the analyzed proteins shown experimentally to form amyloid aggregates or to be located in the core of the mature fibrils formed by these polypeptides. We have defined a parameter called Interface Proximity Index (IPI) to evaluate the degree to which an aggregation-prone region is closer to a given interface than to the rest of the protein surface (see Methods and Figure 1).

Human ß2-Microglobulin

Amyloidosis related to β2-Microglobulin (β2-m) is a common and serious complication in patients on long-term hemodialysis [46]. Two aggregation-prone regions encompassing residues 22–31 and 60–70 were predicted for human β2-m (Figure 2). These regions neatly coincide with two secondary structure elements in β2-m: β-strand 2, formed by residues 21–31, and β-strand 6, formed by residues 61–71. Interestingly, most of the residues in these two regions appear to be solvent accessible (Table 1). In agreement with the prediction, the fragments 21–31 and 21–41 of β2-m self-assemble into fibrillar structures [47]. Also, a peptide corresponding to residues 59–79 and its shorter version 59–71 both form amyloid fibrils [48].

A main interaction cluster is predicted for human β2-m (Figure 2A). It involves Y26 and G29 in β-strand 2, residues H31-S33 in the loop connecting β-strands 2 and 3, residues D53-W60 in β-strand 5 and the adjacent loop, and finally, residues F62 and L63 in β-strand 6. Overall, 62% of the residues in regions with high aggregation propensity are less than 3 Å from predicted "hot spots" of interaction (Table 1), and 25% overlap with them. Specifically, residues at positions 26, 29, 31, 60, 62, and 63 are predicted to be important both for binding and for aggregation.

Class I major-histocompatibility-complex (MHC) molecules (HLA molecules in humans) are ternary complexes of β2-m, an MHC heavy chain, and a bound peptide [49]. The crystal structures of several of these complexes have been solved, providing a benchmark to evaluate the accuracy of the predicted interface. In HLA-A-class molecules, the interface of β2-m and the HLA heavy chain is well conserved [50] and typically comprises 16 β2-m residues: K6, Q8, 10Y, 11S, 12R, N24, Y26, H31, D53, S55, F56, W60, F62, Y63, D98 and M99. This includes 8 of the 15 interacting residues predicted for β2-m. Residues 24, 26, and 31 map to the first aggregation-prone region of β2-m, and residues 60, 62, and 63 map to the second one. Taking as an example the structure of one such HLA-A complex (PDB ID: 1DUZ) [51], 85% of the residues in β2-m aggregating regions are less than 3 Å from the interface in the complex (Table 1 and Figures 2B, 2D and 2E). The IPI values confirm that these regions are preferentially located close to the interface of the complex (Table 1 and Figure 1A).

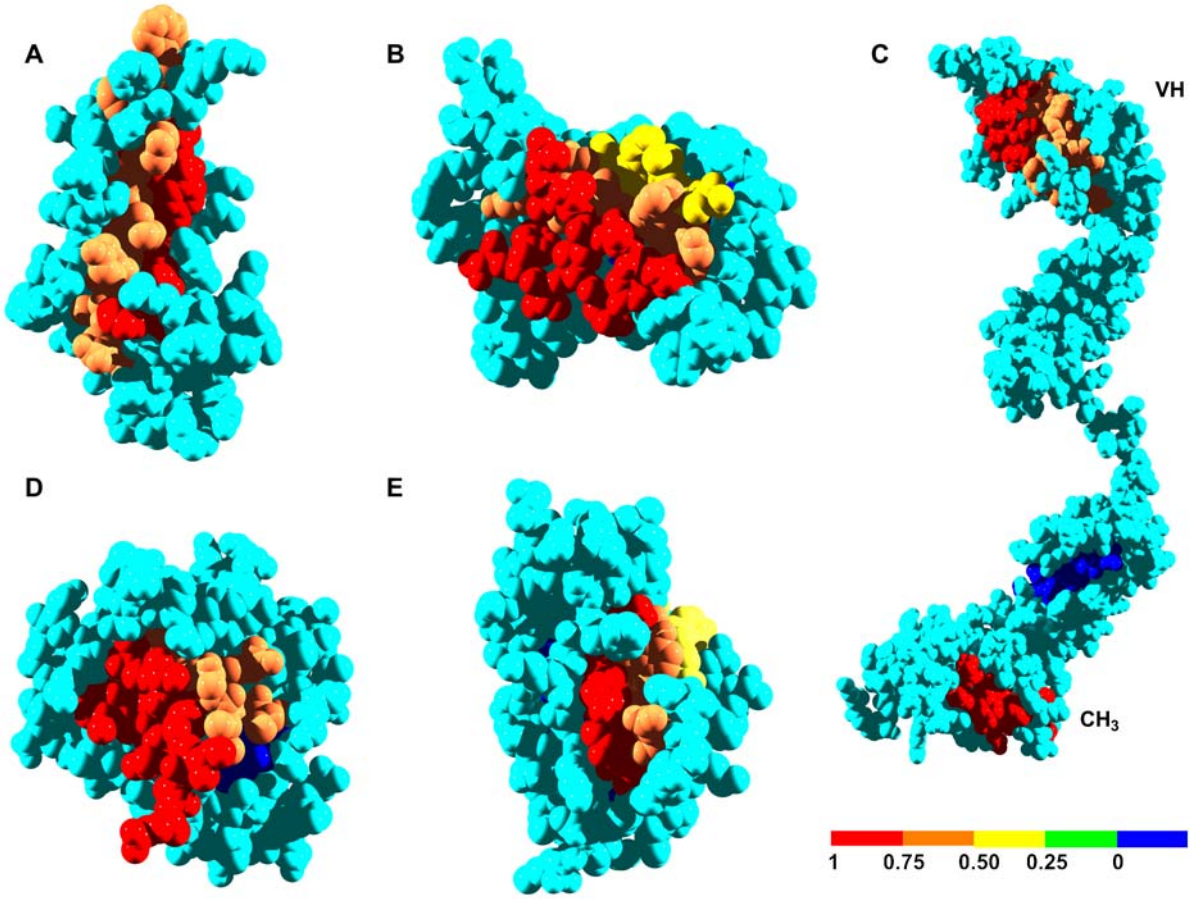


Figure 1. Interface Proximity Index (IPI) of aggregation-prone regions in human globular amyloidogenic proteins. Aggregation-prone regions are coloured according to their IPI values (see the scale). A) β2-microglobulin, B) transthyretin, C) immunoglobulin G heavy chain, D) SOD1 and E) immunoglobulin light chain variable domain.
doi:10.1371/journal.pcbi.1000476.g001

Inside the cell, β2-m associates with the non-classical HLA class I molecule human hemochromatosis protein (HFE) [52]. Hereditary hemochromatosis is a genetic disorder characterized by defects in iron metabolism and associated with mutations in the HFE gene [53]. Some of these mutations prevent the binding of HFE to β2-m. There are 18 β2-m residues at the HFE/β2-m complex interface, according to its crystal structure (PDB ID: 1A6Z) [54] : I1, Q8, 10Y, 11S, 12R, N24, Y26, H31, D53, L54, S55, F56, W60, Y63, F62, L65, D98, and M99, including 9 of the 15 predicted interaction sites. Residues 24, 26, and 31 correspond to the first aggregation-prone region of β2-m and residues 60, 63, 62, and 65 to the second. Another significant feature of this complex is that 76% of the residues in regions with high aggregation potential are close to the interface with β2-m (Table 1 and Figure 2C). Therefore, the docking of the HLA heavy chain and HFE molecules on top of β2-m covers most of the residues in aggregation-prone regions because they are close to the interaction sites, as illustrated by their high IPIs (Table 1, Figure 1A and Figures 2F, 2G).

Aggregation of β2-m under physiological conditions is thought to be initiated by a cis-trans prolyl isomerization of the H31-P32 peptide bond [22]. The transition promotes repositioning of the hydrophobic side chains of F30, L54, F56, W60, F62, and Y63 as shown in the structures of the P32A and P32G mutants [55,56]. Interestingly enough, all of these residues map in an aggregation-prone segment and/or at the interface. Although speculative, it is tempting to propose that conditions that promote the dissociation of β2-m complexes with the above proteins or related ones may uncover this region and facilitate its fluctuation towards amyloidogenic conformations. In fact, *in vivo*, β2-m is continuously shed from the HLA molecules in the cell surface into the serum and transported to the kidneys where it is eliminated. Renal failure increases the levels of circulating β2-m more than 50-fold and promotes its self-assembly and conversion into amyloid fibrils [57]. Consequently, dissociation of β2-m from the class I HLA complex effectively constitutes a critical initial step in its aggregation into amyloid fibrils.

Because the β2-m regions likely to be involved in aggregation are already located in preformed β-strands, local fluctuations may allow anomalous intermolecular interactions between these preformed elements, leading to the formation of an aggregated β-sheet structure without extensive unfolding. In this context, the formation of β2-m complexes both inside the cell and on the cell surface might play a protective role against β2-m aggregation, either by reducing conformational fluctuations or by preventing the exposure of dangerous amyloidogenic regions, or both.

Human Transthyretin

Transthyretin (TTR) constitutes the fibrillar protein found in familial amyloidotic polyneuropathy (FAP), familial amyloidotic cardiomyopathy, and central nervous system amyloidosis. Around 100 different TTR mutations have been reported, many of which

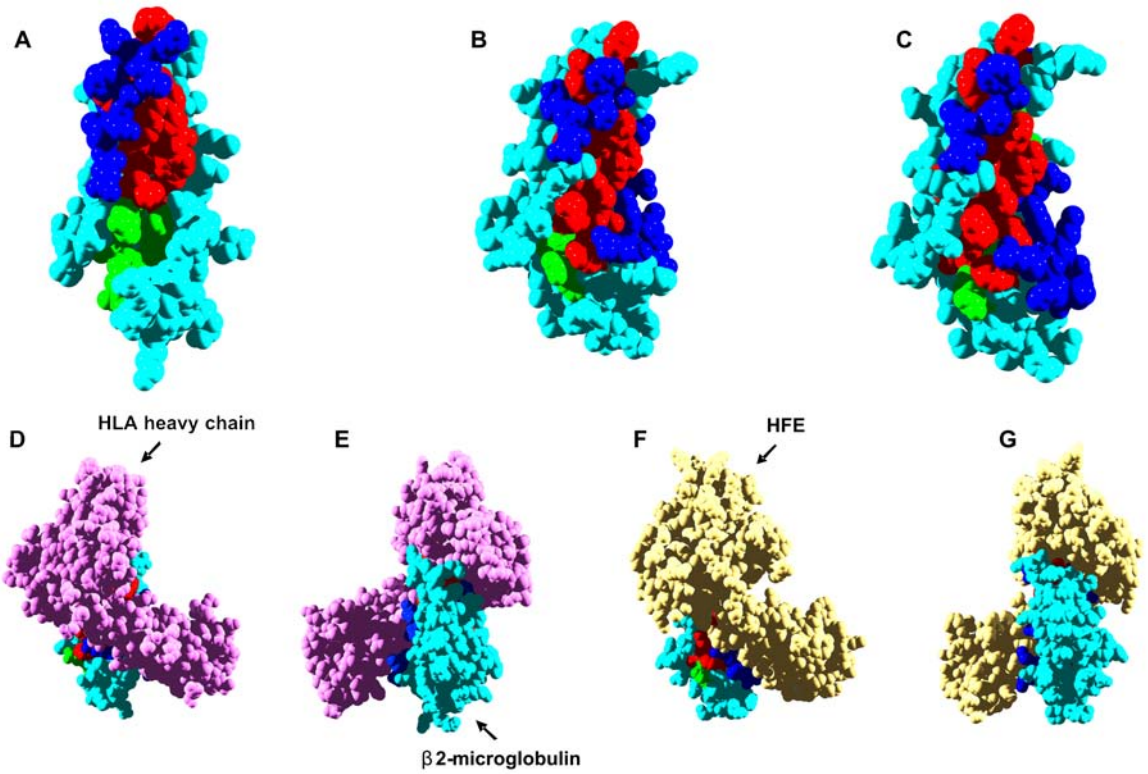


Figure 2. Aggregation and interaction regions in human ß2-microglobulin. In all panels, β2-microglobulin aggregation-prone residues at less and more than 3 Å from interaction sites are shown in red and green, respectively. Interface residues not included in aggregation-prone regions are shown in dark blue. Rest of residues are shown in light blue. A) The predicted interaction surface for monomeric β2-microglobulin is used for calculation. B) The interface between β2-microglobulin and HLA heavy chain is used for calculation (PDB ID:1DUZ). C) The interface between β2-microglobulin and HFE is used for calculation (PDB ID:1A6Z). D and E) Front (same orientation that in B) and back view of the β2-microglobulin/HLA heavy chain complex. F and G) Front (same orientation that in C) and back view of the β2-microglobulin/HFE complex.
doi:10.1371/journal.pcbi.1000476.g002

are amyloidogenic [58]. Native TTR is a homotetramer. Five aggregation-prone regions are predicted for the TTR monomer. They encompass residues 11–19, 26–34, 92–96, 105–112, and 115–121. In this case, the aggregation-prone sequences appear to coincide precisely with preformed β-sheet structures: A β-strand (11–19), B β-strand (26–36), F β-strand (91–97), G β-strand (105–112), and H β-strand (115–121). In concordance with the prediction, peptides 10–20 and 105–115, which map in the first and fourth aggregation-prone regions, have been shown to assemble into amyloid fibrils [59,60].

A single interaction patch is predicted for the TTR monomer (Figure 3A). It involves 19 residues located in the A β-strand (L17, A19), in the loop between the A and B β-strands (V20–S23), in the α-helix (L82), in the loop between the helix and the F β-strand (S85-F87), in the F β-strand (E92), in the G and H β-strands, and in the loop between the G and H β-strands (L110, S112-T118). TTR is a dimer of dimers. In the dimers formed by the A and B or the C and D chains, the predicted clusters are contiguous, forming a large and continuous interaction patch. Of the residues in aggregation-prone regions in TTR, 41% are within 3 Å of predicted interaction sites (Table 1). With the exception of the I26-R34 fragment, all the regions with high aggregation propensity are located close to the predicted interface, and 30% of the residues in these segments overlap with predicted interaction sites. Residues 17, 19, 92, 110, 112, and the stretch 115–118 are predicted to be important both for aggregation and interaction events.

The crystal structure of the TTR tetramer (PDB ID: 1TTA) [61] reveals that the real interfaces between the four individual TTR chains involve residues L17, A19-S23, F87-E89, E92, V94-T96, Y105, L110, and S112-V122. In good agreement with the prediction, the interfaces include 16 of the 19 predicted interacting residues. Residues 17 and 19 map to the first aggregation-prone region, residues 92 and 94–96 to the third one, and residues 110 and 112–122 to the fourth and fifth stretches. Significantly, if we exclude the I26-R34 region (IPI<0), 90% of the residues in aggregating regions are close to the two interfaces of the TTR tetramer as confirmed by their overall high IPIs (Table 1, Figure 1B and Figures 3B, 3C). Accordingly, although these regions are mostly accessible to solvent in the monomer, they become protected in the native quaternary structure of TTR by the interaction of the TTR subunits (Figure 3D).

Dissociation of the TTR tetramer has been reported as a prerequisite for amyloidosis. The tetrameric structure dissociates into AB and CD dimers, but they are unstable in the absence of additional quaternary interactions, explaining why TTR exists in a primarily tetramer-monomer equilibrium [62]. The crystal structures of more than 10 FAP-related variants have been solved, showing that the mutants are essentially identical in tertiary and quaternary structure to the wild-type protein, precluding the presence of preformed conformational defects in the amyloidogenic mutants [63]. However, FAP-associated mutants are destabilized even when tetrameric. This destabilization favors tetramer dissociation to the amyloidogenic monomeric intermediate, exposing previously hidden, preformed, aggregation-prone β-strands. In this context, the overlap of interaction and aggregation surfaces in the AB and CD dimers appears to be an effective way

Table 1. Comparison of aggregation predictions and experimental available data for human globular proteins and proximity of aggregation-prone regions to predicted and real interfaces.

Predicted Aggregation segments	Fibril formers	% residues close to predicted Interface[1]	% residues close to real interface[2]	IPI	% solvent accessible residues
ß2-Microglobulin (HLA)					
22–31	21–31	70 (20)	100 (18)	0.88	65
	21–41				
60–70	59–79	54 (24)	73 (30)	0.58	100
	59–71				
ß2-Microglobulin (HFE)					
22–31	21–31	70 (20)	70 (12)	0.83	65
	21–41				
60–70	59–79	54 (24)	81 (31)	0.62	100
	59–71				
Transthyretin					
11–19	10–20	44 (22)	77 (36)	0.46	66
26–34	-	0 (53)	0 (68)	<0	66
92–96	-	40 (42)	100 (0)	1	100
105–112	105–115	62 (30)	100 (40)	0.6	100
115–121	-	62 (32)	100 (0)	1	100
SOD1					
4–8	-	0 (13)	100 (3)	0.97	66
100–106	-	43 (12)	0 (19)	<0	66
111–120	-	70 (18)	50 (14)	0.72	100
146–153	-	87 (21)	87 (2)	0.97	100
Lysozyme					
25–33	-	33	0 (25)	<0	44
57–66	-	90	40 (32)	0.20	60
76–84	26–123	56	56 (25)	0.55	88
108–114	-	86	0 (46)	<0	100
Immunoglobulin (LC)					
19–23	-	-	0 (38)	<0	71
31–38	-	-	89 (24)	0.71	75
46–51	-	-	50 (30)	0.4	71
71–78	-	-	0 (43)	<0	62
84–89	-	-	83 (18)	0.78	66
Immunoglobulin (HC)					
29–38	-	-	50 (20)	0.60	80
45–52	-	-	75 (25)	0.67	82
87–93	-	-	57 (24)	0.58	57
100–106	-	-	100 (16)	0.84	100
275–281	-	-	0 (31)	<0	71
289–299	-	-	0 (52)	<0	100
322–331	-	-	0 (37)	<0	80
390–396	-	-	63 (13)	0.79	57
435–442	-	-	100 (16)	0.84	75

[1]Percentage of residues in the aggregation-prone region at less than 3 Å from a protein predicted interaction residue.

[2]Percentage of residues in the aggregation-prone region at less than 3 Å from a residue located at the interface of the following complexes: β2-microglubulin in complex with HLA heavy chain [1DUZ] and with HFE [1A6Z]. Native tetrameric structure of transthyretin (PDB code 1TTA). Dimeric structure of SOD1 (PDB code 2C9V). Lysozyme in complex with a camelid antibody (PDB code 1OP9). Dimeric structure of Immunoglobulin LC variable domain (PDB code 2Q20). HCs and LCs of a IgG1 human immunoglobulin (PDB code 1HZH).

[1,2]In brackets the percentage of residues in the aggregation-prone region close to a random surface of the same size than the considered interface.

doi:10.1371/journal.pcbi.1000476.t001

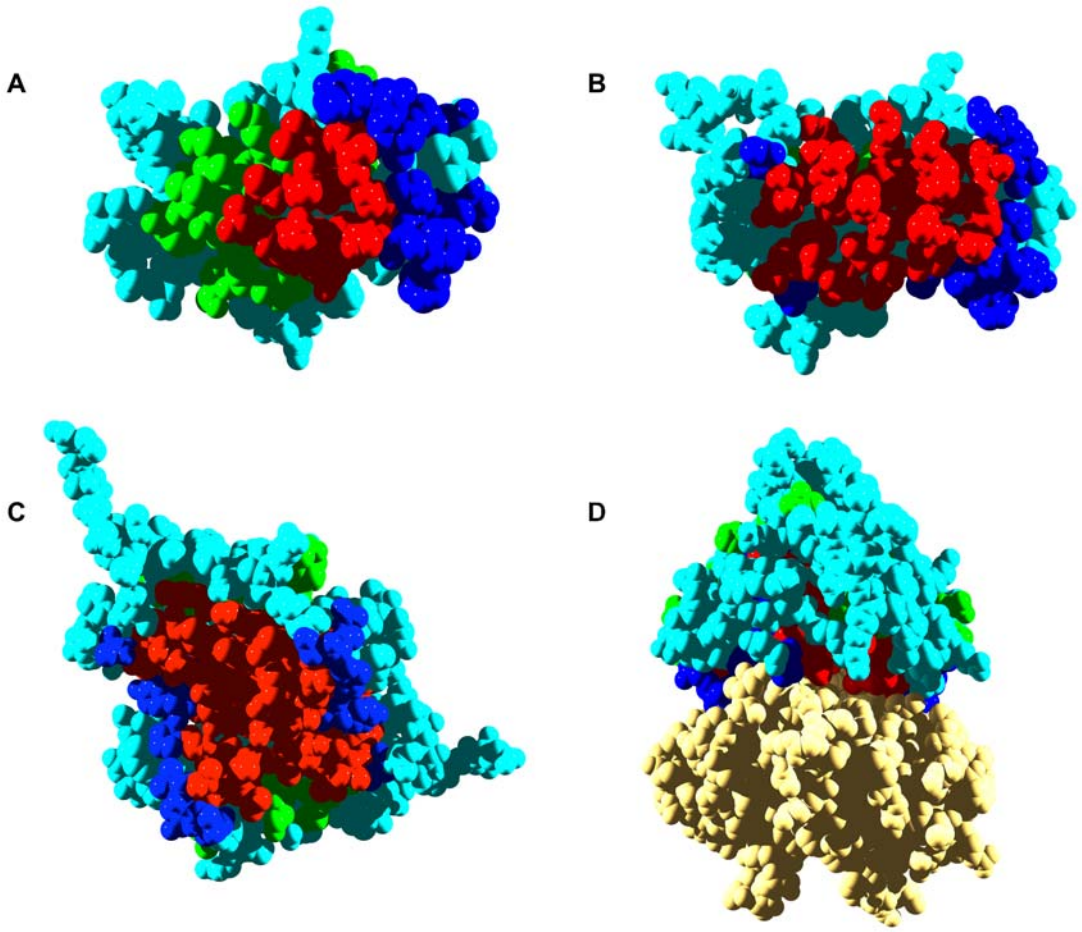


Figure 3. Aggregation and interaction regions in human transthyretin. In all panels, transthyretin (TTR) aggregation-prone residues at less and more than 3 Å from interaction sites are shown in red and green, respectively. Interface residues not included in aggregation-prone regions are shown in dark blue. Rest of residues are shown in light blue. A) The predicted interaction surface of a TTR monomer is used for calculation. B) The interface in the native tetrameric structure of TTR is used for calculation (PDB ID:1TTA). C) Dimer of TTR. D) TTR native tetrameric structure. The first dimer is twisted 90° relative to C, the second one is shown in yellow.
doi:10.1371/journal.pcbi.1000476.g003

to prevent TTR amyloidogenesis in physiological conditions. The success of this strategy is best exemplified by the behavior of the T119M TTR mutant. The presence of the T119M allele alleviates the effect of the aggressive V30M amyloidogenic mutation in patients carrying these two variants. It has been shown that heterotetramers that incorporate T119M subunits are more stable, dissociate at lower rates, and accordingly are less amyloidogenic [64].

Human Copper-Zinc Superoxide Dismutase

Familial amyotrophic lateral sclerosis (fALS) is characterized by the presence of Copper-Zinc Superoxide Dismutase (SOD1) inclusions in spinal cords [65]. Native SOD1 is a homodimer. The SOD1 monomer displays four regions with high aggregation potential. They encompass residues 4–8 in ß-strand 1, 100–106 and 111–120 in ß-strands 6 and 7 and the loop connecting them, and residues 146–153 in ß-strand 8.

A total of 14 residues are predicted to be at the interface of the SOD1 monomer (Figure 4A). They correspond to E21, W32, G33, S105, S107, G108, H110, C111, I113-R115, G147, V148, and I151. Of the residues in aggregation-prone regions in SOD1, 61% are less than 3 Å from predicted interaction sites (Table 1), and 25% of them overlap the predicted interaction sites. In particular, residues 105, 111, 113, 114, 115, 147, 148, and 151 are predicted to be involved in both binding and aggregation.

According to the crystal structure of the SOD1 dimer (PDB ID: 2C9V) [66], the real interface between the two SOD1 subunits involves residues V5, V7, F50-T54, I113-R115, V148, and G150-Q153 (Figure 4B). Therefore, the interaction prediction is poor for the N-terminal part of SOD-1 but accurate for residues in the C-terminal region. Residues V5 and V7 are part of the first aggregation-prone region, S105 part of the third one, I113-R115 part of the fourth stretch, and V148 and G150-Q153 part of the last one. All the residues in the first and last aggregation-prone segments as well as residues C111-T116 are close to the dimer interface (Table 1). Accordingly, except for the 100–106 stretch (IPI<0), all the regions with high aggregation propensity in SOD display high IPIs (Table 1 and Figure 1D). Three out of the four cysteine residues in each SOD1 monomer (6, 111, and 146) are in those sequence stretches. C6 and C111 are present in the form of free cysteines whereas C146 forms a disulfide bond with C57. All of these regions are accessible to solvent in the monomeric form but become partially or totally protected upon dimer association (Figure 4C and 4D).

FALS has been shown to be associated with more than 100 different SOD1 mutations, which are scattered throughout the

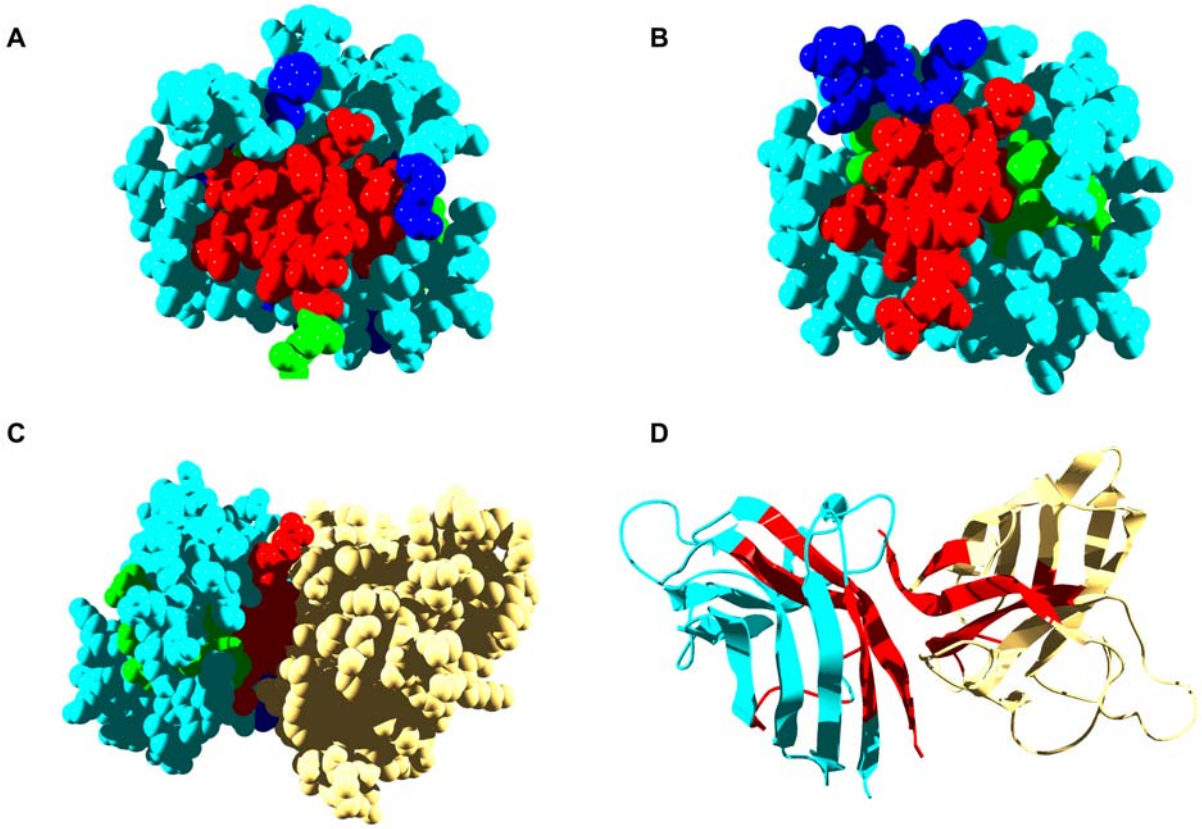


Figure 4. Aggregation and interaction regions in human SOD1. In panels A, B and C SOD1 aggregation-prone residues at less and more than 3 Å from interaction sites are shown in red and green, respectively. Interface residues not included in aggregation-prone regions are shown in dark blue. Rest of residues are shown in light blue. A) The predicted interaction surface of a SOD1 monomer is used for calculation. B) The interface in the native dimeric structure of SOD1 is used for calculation (PDB ID:2C9V). C) Native dimer of SOD1, the second monomer is shown in yellow. D) Ribbon representation of the SOD1 dimer, predicted aggregation-prone regions are shown in red.
doi:10.1371/journal.pcbi.1000476.g004

three-dimensional structure [67]. Among them, the A4V mutation has received special attention because it results in a rapidly progressing form of fALS [68]. Animal models suggest that the pathogenicity of the A4V SOD1 arises from an increased propensity to aggregate, forming amyloid fibrils or pores [69]. A4 is near the dimer interface and maps in the first aggregation-prone region. Hasnain and co-workers solved the crystal structures of dimeric forms of A4V and another FALS mutant, I113T [70]. I113 is also at the interface, in the third aggregation-prone region. Both variants display the same monomer fold and active-site geometry as WT, but their interfaces are destabilized. Ray and Lansbury have shown that a covalent link between the two A4V SOD1 subunits abolishes aggregation, suggesting that the monomer is an obligate intermediate along the aggregation pathway [71]. Other studies also support the idea that monomerization leads directly to aggregation and fibrilization [72]. However, other lines of evidence suggest that the cytotoxic properties of SOD1 are triggered by an incorrect connection of its cysteine residues. In support of this view, the toxicity of recombinant SOD1 in cultured cells is lost upon mutational removal of C6 and C111 [11], and nucleation of the aggregation reaction requires the presence of cysteine thiolates at both positions 57 and 146 [72]. In any case, it appears that the interface plays a protective role against aggregation in SOD1, by preventing the direct assembly of pre-formed and exposed aggregation-prone regions in the monomer, by stabilizing the monomer against conformational fluctuations that might expose amyloidogenic sequences, or by preventing the exposure and reshuffling of cysteine residues. Based on these observations, it has been proposed that the stabilization of the SOD1 dimer interface could become an effective approach to fight against fALS [71].

Human Immunoglobulins

The light chains (LCs) of immunoglobulins have been implicated in the pathogenesis of amyloidosis in patients with monoclonal B-cell proliferative disorders (AL amyloidosis) [73]. When immunoglobulin molecules are secreted, two heavy chains (HCs) usually pair with two LCs to create a heterotetramer. Occasionally, free LCs are secreted, and these LCs can form homodimers. LC dimers can be innocuous, but they can also aggregate into pathogenic species. We have analyzed the aggregation propensity and interfaces of a non-pathogenic LC dimer (PDB ID: 2Q20) [74]. Five aggregation-prone regions are detected, encompassing residues 19–23, 31–38, 46–51, 71–78, and 84–89 located in the ß3, ß4 ß5, ß9, and ß10 strands, respectively (Table 1). The interface of the dimer involves 13 residues: D34, Y36, Q38, K42-P44, L46, E55, Y87, Q89, Y91, Y96, and F98. According to their IPIs, the second and fifth stretch are located preferentially at the interface of the complex, with 89% and 83% of their residues less than 3 Å from the interface, respectively (Table 1, Figure 1E and Figures 5A, 5B). It is important to note that both stretches map in preformed ß strands.

AL is distinct from other types of amyloidosis in that hypervariability yields a different set of mutations in each patient. Ramirez-Alvarado and co-workers have characterized an LC

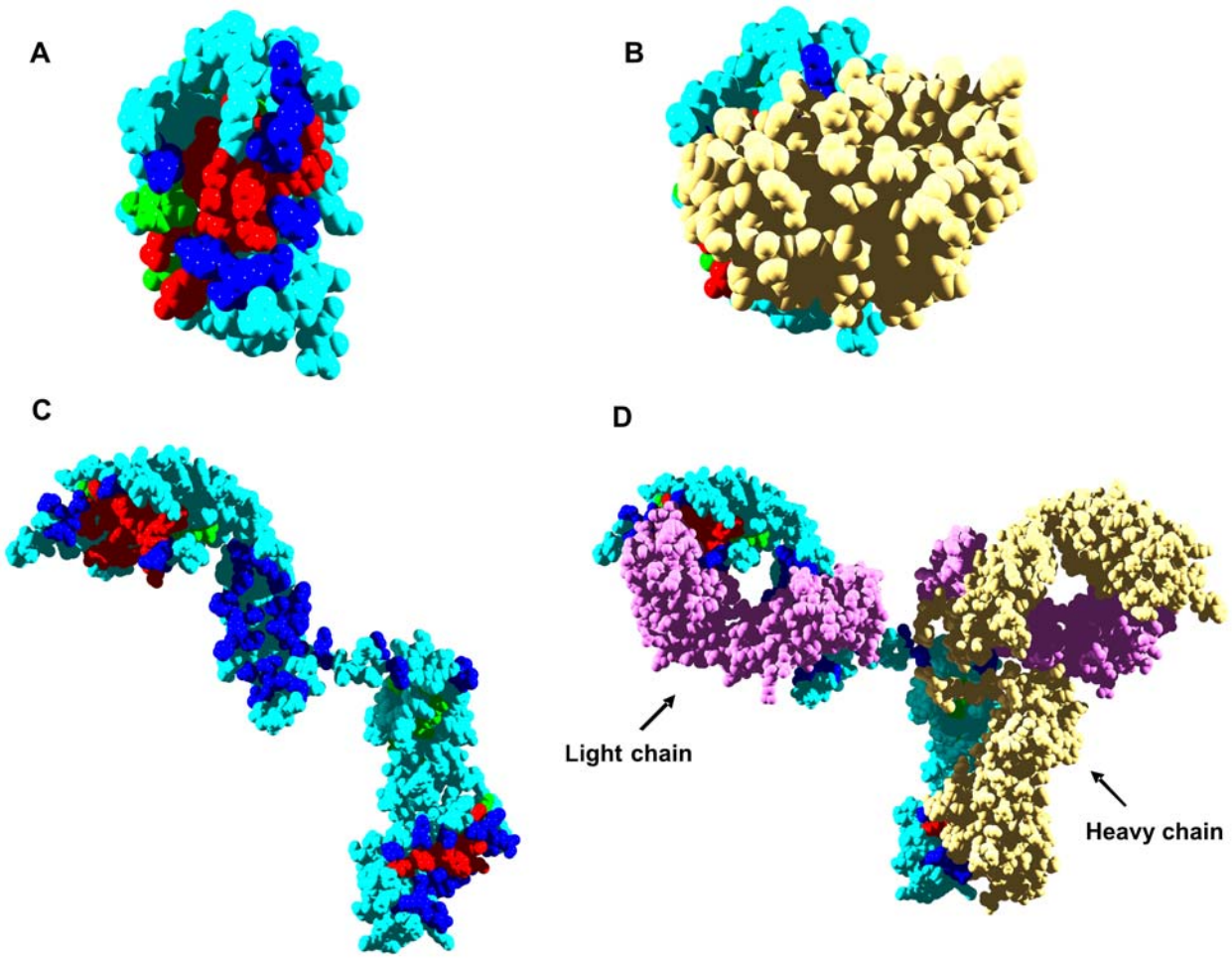


Figure 5. Aggregation and interaction regions in human immunoglobulins. In all panels, immunoglobulin (Ig) aggregation-prone residues at less and more than 3 Å from interaction sites are shown in red and green, respectively. Interface residues not included in aggregation-prone regions are shown in dark blue. Rest of residues are shown in light blue. A) The interface in the native structure of Ig light chain variable domain (LC) is used for calculation (PDB ID: 2Q20). B) Native homodimer of Ig LC, the second monomer is shown in yellow. C) The interface in the native structure of IgG heterotetramer is used for calculation and the Ig heavy chain (HC) represented (PDB ID: 1HZH). D) Native IgG heterotetramer. Ig LCs and the second Ig HC are indicated.
doi:10.1371/journal.pcbi.1000476.g005

dimer isolated from an AL patient [74]. The pathogenic protein differs from its germline in seven residues. Only three changes are non-conservative, and all of them are located at the dimer interface: N34I, K42Q, and Y87H. The N34I and Y87H mutations occur precisely in the second and fifth aggregation prone regions in the protein. Ramirez-Alvarado and co-workers found that the mutant dimer has an interface that is rotated 90° from the canonical LC interface. The altered interface was accompanied by decreased thermodynamic stability of the dimer and accelerated fibril formation. This might result from the exposure and self-assembly of the above preformed aggregation-prone ß segments upon dimer destabilization or dissociation. Interestingly, the restorative mutation H87Y suffices to regain thermodynamic stability, delay amyloid formation, and restore the canonical dimer interface, illustrating a delicate balance between native and aberrant protein self-assembly.

Although AL is more frequent, in some systemic amyloidosis the amyloid deposits consist of an unusual form of IgG1 heavy chain (HC) [75]. The amyloid protein contains the complete heavy-chain variable (VH) domain contiguous to the third constant region (CH_3) due to the total absence of the first (CH_1) hinge and second (CH_2) heavy-chain constant regions [75].

Using the structure of a complete human IgG1 antibody [76] as a model (PDB ID: 1HZH), we detected nine aggregation-prone regions in the heavy chain (Table 1). Four of the aggregation-prone regions are in the VH domain (29–38, 45–52, 87–93, and 100–106), three in the CH_2 domain (275–281, 289–299, and 322–331), and two in the CH_3 domain (390–397 and 435–442). Analysis of the structure of the oligomeric form of the antibody reveals that only the regions in the VH and CH_3 domains of the heavy chain display high IPI values and therefore are adjacent to the interface in the native heterotetramer (Table 1, Figure 1C and Figures 5C, 5D). The truncated, pathogenic form of the IgG is found in monomeric form in urine, indicating that either it cannot associate or it dissociates from the light and heavy chains that block the exposure of the detected aggregating regions in a normal heterotetrameric IgG molecule. These sequence stretches are located in preformed ß strands and are ready for self-assembly reactions that might result in the observed amyloid deposits.

Protein Binding Prevents Aggregation: Human Lysozyme and Aß42

Human lysozyme forms amyloid fibrils in individuals suffering from nonneuropathic systemic amyloidosis. The disease is always

associated with non-conservative point mutations in the lysozyme gene [77]. Four aggregation-prone regions were detected in human lysozyme, corresponding to residues 25–33, 57–66, 76–84, and 108–114. The first region maps in helix B, the second and third in the loop of the β-domain, and the last one around the short helix D (Table 1). In good agreement with the predictions, recent experimental data shows that the region comprising residues 26–123 is preferentially protected from proteolysis once it is incorporated into lysozyme amyloid fibrils [78].

Two different interaction clusters are predicted for human lysozyme (Figure 6A), one in the α-domain and the other in the ß-domain. The first involves residues in the loop of the β-domain: N60, R62-W64, N66, A73-N75, A76, and H78. The second cluster is located in helix C and around helix D and corresponds to residues A94, K97, R98, R107-W109, and W112. Residues K33 and W34 in helix B are also predicted to be involved in protein-protein interactions. Overall, 66% of the residues in regions with high aggregation propensity are less than 3 Å from predicted interaction sites, and 31% overlap with them. Residues 33, 60, 62–64, 66, 76, 78, 108, 109, and 112 might be implicated in both binding and aggregation reactions. Interestingly, residues I56, F57, W64, and D67, which are mutated in the four known single-residue familial variants associated with lysozyme amyloidosis, are comprised of or very close to protein segments with high aggregation propensity and/or interaction sites.

The mechanism of lysozyme aggregation under physiological conditions probably involves thermal fluctuations that transiently expose amyloidogenic regions [22]. These transitions are rare in the wild type protein, but they are more frequent in mutated forms related to amyloidosis. It has been suggested that residues 36–102 in the β-domain and helix C can unfold while the rest of the α-domain maintains a native-like conformation [9]. In particular, residues 78–80 have been proposed to have a high aggregation propensity and the lowest structural protection, and therefore the highest propensity to initiate aggregation [79]. This sequence includes predicted interacting residues in the loop of the ß-domain and also overlaps with the predicted 76–84 amyloidogenic region.

A single-domain fragment of a camelid antibody has been shown to inhibit the *in vitro* aggregation of the D67H amyloidogenic lysozyme variant [80]. The antibody epitope includes neither the site of mutation nor most of the protein region destabilized by the mutation; therefore it was suggested that the binding of the antibody prevents aggregation by restoring the structural cooperativity of the mutant protein through the transmission of long-range conformational effects [80]. The structure of the antibody-lysozyme complex (PDB ID: 1OP9) reveals that the epitope consists of 14 residues of the lysozyme molecule and encompasses residues located in the loop between the A and B helices in the α-domain (L15, G16, Y20), in the long loop within the ß-domain (A76, C77, H78, L79), and in the C-helix (A90, D91, A94, C95, K97, R98, R101) (Figure 6B). The epitope includes interaction residues in the first and second predicted clusters. Also, the residues in the loop of the ß-domain coincide with the 76–84 aggregation-prone region. Therefore, an

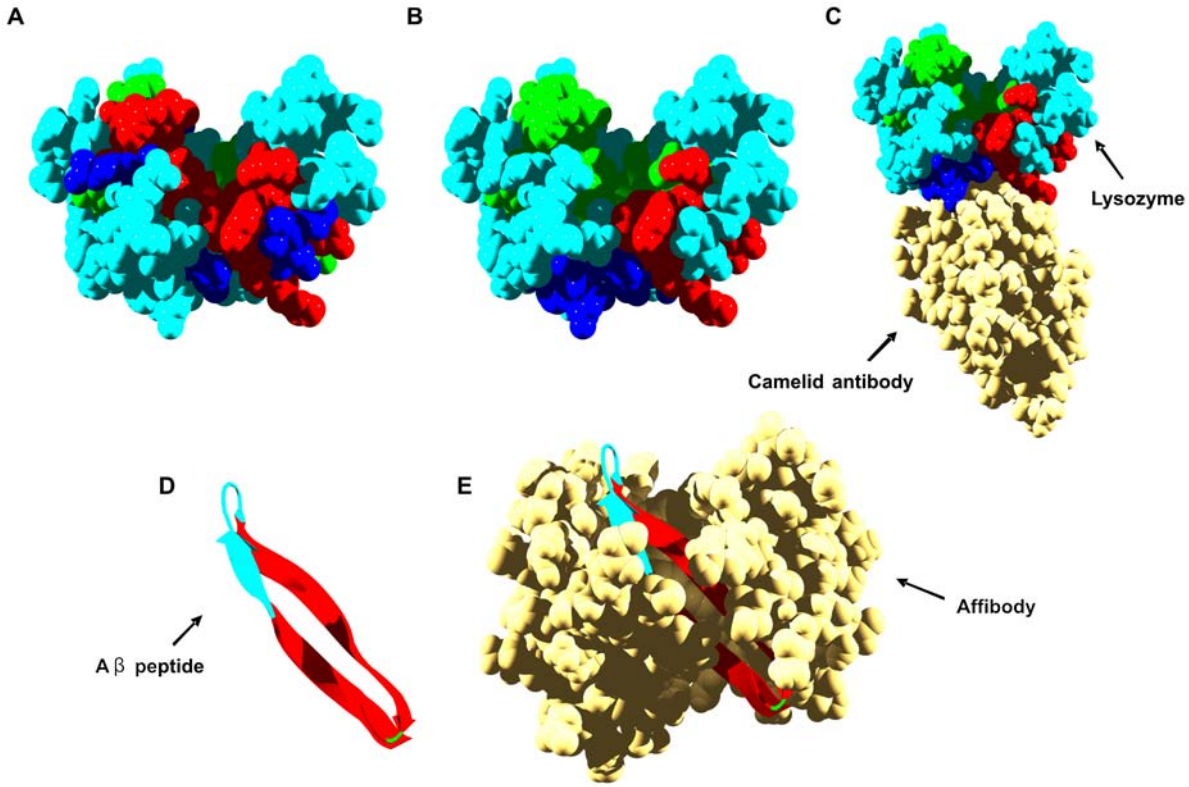


Figure 6. New interfaces at human lysozyme and Aß peptide aggregation-prone regions. In all panels, aggregation-prone residues at less and more than 3 Å from interaction sites are shown in red and green, respectively. Interface residues not included in aggregation-prone regions are shown in dark blue. Rest of residues are shown in light blue. A) The predicted interaction surface of lysozyme is used for calculation. B) The interface between lysozyme and a camelid antibody is used for calculation (PDB ID: 1OP9). C) Lysozyme complex with a camelid antibody. D) Ribbon representation of Aß peptide. The interface between the peptide and a designed affibody is used for calculation (PDB ID: 2OTK). E) Aß peptide bound to a designed affibody.
doi:10.1371/journal.pcbi.1000476.g006

alternative explanation for the protective action of the antibody could be that by docking on top of interaction clusters, it impedes the conformational fluctuation and exposure of the amyloidogenic region around residues 70–80 (Figure 6C).

A nice example illustrating how new binding interfaces can effectively inhibit amyloid formation has been recently reported for the Alzheimer's Aβ peptide. Two aggregation-prone regions comprising residues 16–21 and 29–40 are consistently predicted for Aβ (Figure 6D). The prediction is in excellent agreement with the experimental data in the literature indicating that these regions constitute the core of the Aβ fibrils [81]. Härd and co-workers have used the Z domain derived from staphylococcal protein A to evolve variants of this domain able to bind to Aβ with nanomolar affinity and abolish its aggregation (affibodies) [82]. The solution structure of one of these complexes illustrates how the affibody's protective effect is exerted by creating a new, continuous interface with Aβ that buries its two aggregation-prone regions within a large hydrophobic tunnel-like cavity (Figure 6E).

Non-Amyloidogenic Monomeric Proteins

An important question to address is whether predicted interaction interfaces and aggregation-prone regions also coincide in monomeric and soluble proteins. Therefore, we have analyzed the predicted properties of four well-characterized soluble proteins: myoglobin, maltose binding protein, thioredoxin, and ubiquitin.

Human myoglobin is a compact protein not related to disease. Although after long exposure to high temperatures *in vitro* it unfolds and assembles into amyloid fibrils [15], it is a highly soluble protein in its native α-helical conformation. It displays four regions with high aggregation potential encompassing residues 8–15, 28–33, 67–76, and 110–117. This last segment partially overlaps with the peptide fragment 100–114 found to form amyloid structures *in vitro* [83]. A 12-residue interface is consistently predicted for myoglobin. It consists of residues L40, K42, F43, L89, S92, I99, P100, K102, Y103, I107, L137, and F138. Interestingly enough, only one residue (I111) in the predicted aggregating regions is close to the interface. In addition, its side chain is buried, resulting in a surface where predicted interaction and aggregation regions do not overlap (Figure 7A), a feature that might have evolved to resist aggregation.

Maltose binding protein (MBP) endows fused proteins with increased solubility indicating that it is by itself highly soluble [84]. However, because it is a relatively large protein (370 residues), 10 different aggregation prone regions are predicted, comprising a total of 82 residues. Similarly to the case of myoglobin, although 8 of these residues are close to the predicted interface, comprising residues F92, E153, F156, M321, E322, A324-I329 and W340, their side chains are not significantly exposed to solvent (Figure 7B).

Thioredoxin A (TRX) is another tag used to increase the solubility of recombinant proteins [85]. Three aggregation-prone regions comprising residues 22–27, 29–33, and 49–57 are detected in human TRX. The predicted interaction surface comprises residues T30-I38, D60, V71-T74, and A92. While the first and third aggregation stretches are at more than 3 Å of the predicted interface, the second one overlaps with it. Surprisingly, in contrast to myoglobin and MPB, this region is exposed to solvent (Figure 7C). This suggests that, as discussed in the previous section, it could be involved in protein assembly reactions. In fact, residues C32–C35 in this stretch constitute the consensus CXXC motif in the TRX active site. In agreement with this hypothesis, we found that in the solution structure of human TRX in a mixed disulfide intermediate complex with its target peptide from the transcription factor NF k-B, the second aggregation-prone region in TRX is part of the complex interface [86] (Figure 7D).

The question arises of why TRX does not self-assemble when it is free. It appears that evolution uses negative design to fight against protein deposition by placing amino acids that counteract aggregation at the flanks of protein sequences with high aggregation propensity [45]. These residues are called aggregation gatekeepers [87], and they reduce self-assembly using the repulsive effect of charge (Arg, Lys, Asp and Glu), the entropic penalty on aggregate formation (Arg and Lys), or incompatibility with ß-structure backbone conformation (Pro) [88]. Interestingly, P34 is adjacent in sequence to the TRX 29–33 aggregation prone region. P34 and the two basic, protruding K37 and K39 residues flank this region in the 3D-structure (Figure 7C), which overall would make self-assembly reactions far more difficult.

Ubiquitin is a small, soluble and highly conserved regulatory protein that is ubiquitously expressed in eukaryotes [89]. Three aggregation-prone regions are detected in ubiquitin, including residues 1–8, 42–47, and 67–74 in the ß1, ß3, and ß5 strands, respectively. In this case, the regions of the protein with the highest aggregation propensity overlap significantly with the predicted interaction interface (Figure 7E). This suggests that in principle, this surface is competent for protein assembly reactions. Importantly, it has been shown that ubiquitin binding motifs, such as CUE domains, bind precisely to a surface defined by the ß1, ß3, ß4, and ß5 strands of ubiquitin (Figure 7F) [90], illustrating again how aggregation-prone regions and interaction interfaces tend to overlap. In fact, biochemical and genetic analyses have defined the hydrophobic patch formed by the side chains of L8, I44, and V70 on the surface of ubiquitin as a key determinant for endocytosis and proteosomal degradation [91]. These three residues are located in each of the three aggregation-prone regions predicted for ubiquitin. Why, then, does ubiquitin not self-assemble when it is unbound in solution? An examination of the surface defined by the above ß-strands shows that ubiquitin uses negative design principles to avoid aggregation, placing a large number of positively charged residues on the edge of these strands and adjacent to them (Figure 7E). Upon binding to ubiquitin-binding domains, these basic residues are hidden at the complex interface.

Non-Amyloidogenic Dimeric Proteins

It seems that the spatial coincidence of interfaces and sequences promoting self-assembly is not restricted to amyloidogenic proteins. To further confirm this extent, we analyzed the structure of 25 different eukaryotic proteins shown to form homodimers (Table 2 and Figure 8). As expected, the number of predicted aggregation-prone regions in a protein correlates with its size ($R = 0.88$). All the analyzed proteins present at least one aggregation segment in which half of the residues are closer than 3 Å to the interface, and 96% of them have at least one aggregation region in which >85% of the residues are adjacent to the interface (Table 2 and Figure 8). This supports the idea that the physico-chemical determinants of aggregation and native self-assembly might overlap significantly and is consistent with the observation that in homodimers, identical monomer subunits tend to associate by hydrophobic interactions [92]. After protein synthesis and folding, monomers probably associate rapidly into native homodimers due to the high local concentration of identical polypeptide chains, thus avoiding prolonged exposure of hydrophobic, aggregation-prone regions to solvent. Interestingly, in heterodimers, in which monomers spend a larger part of their lifetime in a non-associated state, the presence of gatekeeper amino acids (Lys, Arg, Glu, Asp, and Pro) at the complex interface

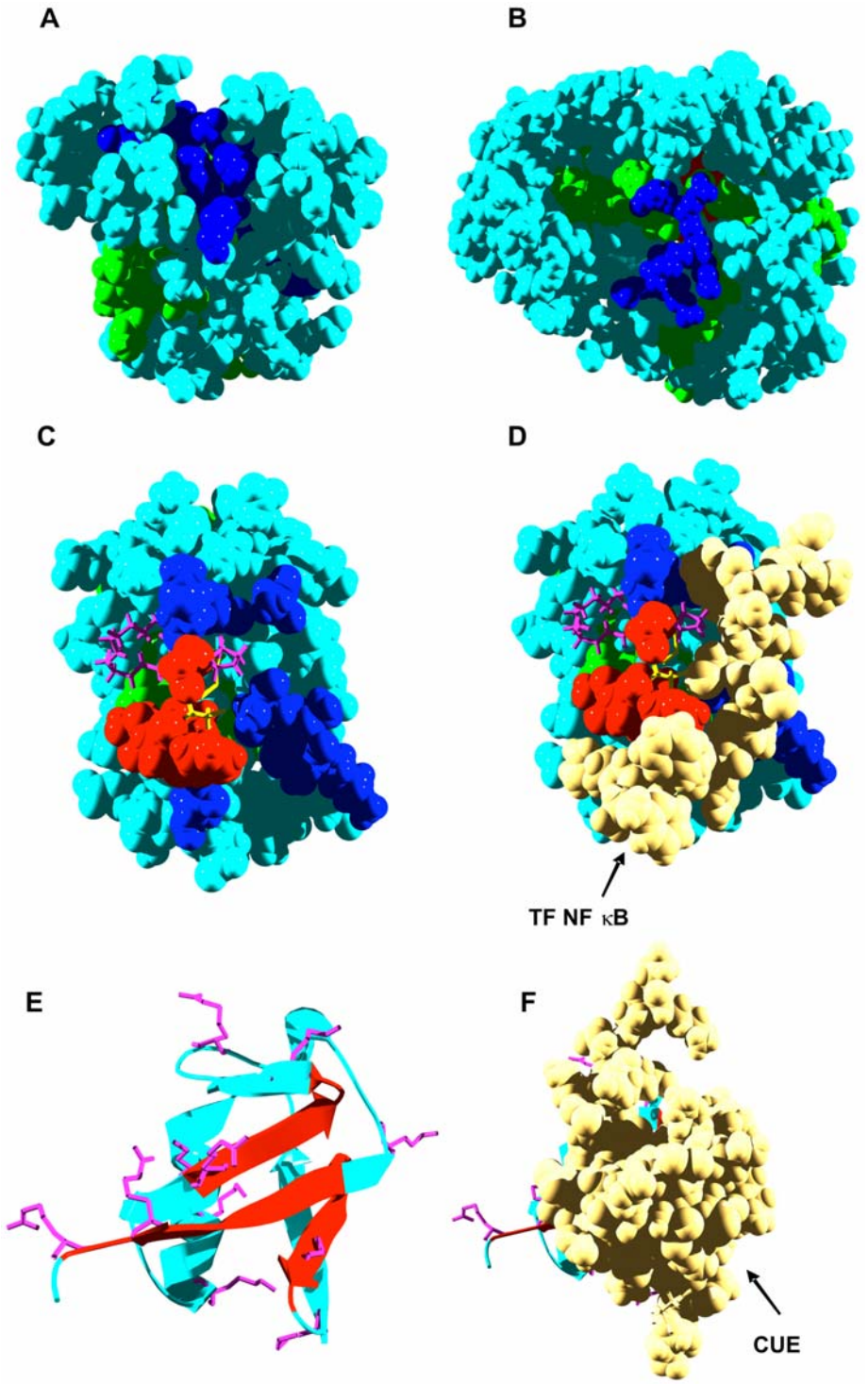


Figure 7. Aggregation and interaction regions in monomeric soluble proteins. In panels A–D, aggregation-prone residues at less and more than 3 Å from interaction sites are shown in red and green, respectively. Interface residues not included in aggregation-prone regions are shown in dark blue. Rest of residues are shown in light blue. In all panels the predicted interaction surface is used for calculation. A) Human myoglobin (PDB ID: 4MBN). B) Maltose Binding Protein (MBP) (PDB ID: 4MBP). C) Human thioredoxin (TRX) (PDB ID: 3TRX). Gatekeeper residues are shown in purple and active cysteines in yellow. D) Same orientation that in C, Human TRX in a mixed disulfide intermediate complex with a peptide from the transcription factor NF kappa B (PDB ID: 1MDI). E) Ribbon representation of human ubiquitin (PDB ID: 1UBQ). Aggregation-prone secondary structures near the interface are shown in red. Basic residues in the vicinity of aggregation-prone regions are shown in purple. F) Same orientation than E). Complex of human ubiquitin with a CUE ubiquitin binding domain (PDB ID: 1OTR).
doi:10.1371/journal.pcbi.1000476.g007

Table 2. Overlapping of aggregation-prone regions and interfaces in non-amyloidogenic eukaryotic homodimers.

PDB	Protein	Source	Length	Aggregation segments	Aggregation segments close to the interface (>50%)[1]	Aggregation segments close to the interface (>85%)[2]
1F17	Dehydrogenase	Homo sapiens	293	9	2	2
1DQT	Antigen	Mus musculus	117	8	4	3
1LR5	Auxin binding protein	Zea mays	160	7	4	1
1KSO	Calcium-binding protein A3	Homo sapiens	93	3	2	2
1EAJ	Coxsackie virus	Homo sapiens	124	4	2	1
1PE0	DJ-1	Homo sapiens	187	7	1	1
1JR8	Erv2 protein mitochondrial	Saccharomyces cerevisiae	105	5	2	2
1F4Q	Grancalcin	Homo sapiens	161	6	3	1
1DQP	Guanidine phosphoribosyltransferase	Giardia lamblia	230	10	3	2
3SDH	Hemoglobin	Scapharca inaequivalvis	145	5	2	2
2HHM	Hydrolase	Homo sapiens	272	11	4	3
8PRK	Inorganic pyrophosphatase	Saccharomyces cerevisiae	282	8	3	2
1QMJ	Lectin	Gallus gallus	132	5	1	1
1M6P	Phosphate receptor	Bos taurus	146	5	2	1
1MNA	Polyketide synthase	Streptomyces venezuelae	276	10	2	1
1F89	Protein YLC351C	Saccharomy cerevisiae	271	11	3	2
1LHP	Pyridoxal kinase	Ovis aries	306	10	3	1
1QR2	Quinone reductase type 2	Homo sapiens	230	9	5	1
3LYN	Sperm lysine	Haliotis fulgens	122	6	3	1
1SCF	Stem cell factor	Homo sapiens	116	4	1	0
1HQO	URE2 protein	Saccharomyces cerevisiae	221	8	3	3
1HSS	Alpha-amylase inhibitor	Triticum aestivum	111	3	2	2
1KIY	Trichodiene synthase	Fusarium sporotrichioides	354	12	4	4
1MI3	Xylose reductase	Candida tenuis	319	6	1	1
1LBQ	Ferrochelatase	Saccharomyces cerevisiae	356	12	3	2

[1]More than 50% of the residues in the aggregation-prone region are at less than 3 Å from a residue located at the interface of the complex.
[2]More than 85% of the residues in the aggregation-prone region are at less than 3 Å from a residue located at the interface of the complex.
doi:10.1371/journal.pcbi.1000476.t002

is much greater than in homodimers [92], probably to prevent self-association between identical monomers.

During the revision of the present work, Vendruscolo and co-workers published a related study in which they used their algorithm Zyggregator to perform an extensive analysis of interfaces in protein-protein complexes [93]. Interestingly enough, they independently concluded that interface regions are more prone to aggregate than other surface regions. Also, in excellent agreement with our analysis on monomeric soluble proteins, they found that charged residues frequently disrupt hydrophobic patterns at interfaces and that regions of negative aggregation propensity tend to surround aggregation-prone regions, which suggests that monomeric and native oligomeric proteins have evolved similar strategies to prevent misassembly. In our study, the analyzed eukaryotic proteins were randomly selected from a dataset of non-redundant homodimers [92], without any previous knowledge of their 3D-structures. Interestingly enough, the aggregation-prone sequences near to the dimer interface are located in α-helices in ~70% of the cases (Figure 8). This is in clear contrast with their location in globular amyloidogenic polypeptides, where they reside mainly in preformed ß-strands. Although the sample is not statistically significant, this observation might suggest that natural selection is acting against the presence of amyloidogenic ß-strands at homodimers interfaces. It is attractive to propose that, as shown here for amyloidogenic proteins, mutations at these protein interfaces and specifically at protective locations might lead to loss of function or toxic phenotypes in a significant number of, yet undescribed, human polypeptides.

Conclusions

In the present work, we have used computational tools to predict aggregation-prone regions and interaction sites in globular proteins related to depositional diseases and non-pathogenic

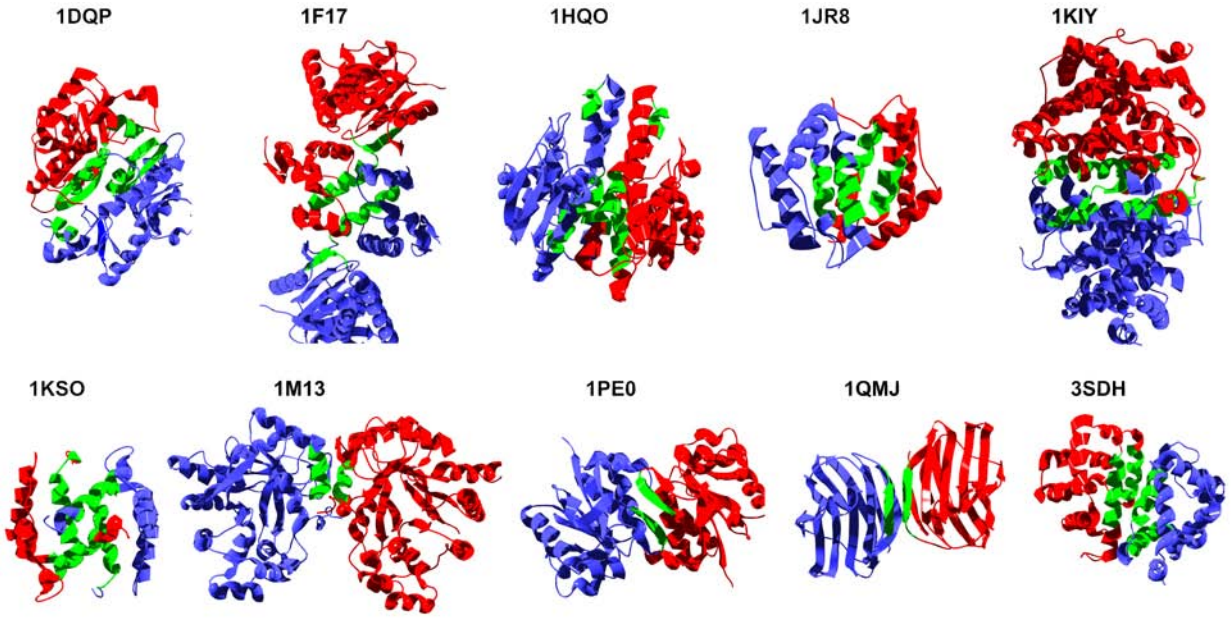


Figure 8. Aggregation-prone regions at the interface of selected homodimeric eukaryotic proteins. Aggregation-prone regions in which more than 85% of the residues are at less than 3 Å from the interface are highlighted in green. The PDB ID is indicated for each dimer (see also Table 2).
doi:10.1371/journal.pcbi.1000476.g008

polypeptides. From the comparison of the predictions with the structural and experimental data, it appears that protein-protein interaction surfaces and regions with high aggregation propensity overlap significantly in the quaternary structure of proteins.

The proximity and coincidence of protein-protein interfaces and aggregation-prone regions suggests that the formation of native complexes and the aggregation of their monomeric subunits probably compete in the cell. This implies that the molecular machinery that performs the vast array of cellular functions and the aggregates that might interfere with these functions promoting cell stress or even cell death are sustained by similar molecular contacts. It is likely that the specificity of native protein interfaces in protein complexes has evolved to minimize anomalous interactions and therefore detrimental protein aggregation reactions. In this sense, Vendruscolo and co-workers have recently identified disulfide bonds and salt bridges as specific interactions that can stabilize aggregation-prone interfaces in their native conformations in oligomeric proteins [93]. However, the balance between functional and aberrant self-assembly appears to be so delicate that point mutations that affect the interface or the stability of the complex, promoting a higher dissociation rate, usually lead to the formation of toxic aggregates, either through direct assembly of newly exposed aggregation-prone regions or by local unfolding of protein segments previously stabilized in the native structure of the complex.

Overall, the present analysis provides a rational to understand how globular proteins aggregate under physiological conditions, where they posses an initially folded and cooperatively sustained conformation and extensive denaturation is not expected to occur. The data strongly suggest that the stabilization of the interface in multimeric proteins, as in the case of TTR, SOD1, or LC immunoglobulins, and/or the blocking of conformational fluctuations and exposed amyloidogenic regions through the formation of new interfaces with other protein molecules, as in the case of lysozyme or Aß peptide, might be important strategies to delay the onset or slow the progress of conformational diseases caused by globular proteins.

The observed association between the failure to attain a native interface and the build up of harmful aggregates suggests that the range of genetic human diseases which ultimately might originate from the conversion of a soluble globular protein into toxic assemblies could be much larger than previously thought. Approaches combining the prediction of aggregation-prone regions from the linear protein sequence with the analysis of real or predicted protein interfaces in the 3D-structure might provide a means to identify physiologically and therapeutically relevant amyloidogenic sequences in the proteins linked to such disorders.

Methods

Prediction of Aggregation-Prone Regions

Aggregation-prone regions in the studied proteins were predicted using the primary sequence as input and a consensus of the output of four different available methods. The first algorithm we used is TANGO (http://tango.crg.es/). TANGO is based on the physico-chemical principles underlying ß-sheet formation, extended by the assumption that the core regions of an aggregate are fully buried [38]. The second algorithm employed was AGGRESCAN (http://bioinf.uab.es/aggrescan/). AGGRESCAN is based on the use of an aggregation-propensity scale for natural amino acids derived from *in vivo* experiments [40]. The third method, developed by Galzitskaya and co-workers, is based on the use of a packing density scale for natural amino acids and on the assumption that amyloidogenic regions are highly packed in the fibrillar structure [41]. The last approach was developed by Zhang and co-workers (ftp://mdl.ipc.pku.edu.cn/pub/software/pre-amyl/). It uses the microcrystal fibrillar structure of the prion hexapeptide NNQQNY [94] as a template and a residue-based statistical potential to identify amyloidogenic fragments of proteins [43]. All analysis was performed using the default parameters for each employed algorithm. In the present work, a sequence stretch in the analyzed proteins should comprise a minimum of five consecutive residues and be positively predicted

by at least two of the above-mentioned methods to be considered an aggregation-prone region.

Prediction of Protein-Protein Interaction Sites

Interaction residues were predicted using the monomeric three-dimensional crystal structure of each of the studied proteins as input and a consensus of the output of three different algorithms. The first approach used to predict interaction surfaces was the Optimal Docking Area (ODA) method (http://www.molsoft.com/oda), which identifies continuous surface patches with optimal docking desolvation energy based on atomic solvation parameters adjusted for protein-protein docking [32]. Only the top ten ODA hot spots were considered. The second method we used was SHARP^2 (http://www.bioinformatics.sussex.ac.uk/SHARP2). SHARP^2 calculates multiple parameters for overlapping patches of residues on the surface of a protein. It considers the solvation potential, hydrophobicity, accessible surface area, residue interface propensity, planarity, and protrusion. Parameter scores for each patch are combined, and the patch with the highest combined score is predicted as a potential interaction site [31]. The patch size was selected by considering the interacting partner to be an identical protein, and only residues in the best-scoring patch were considered. The last algorithm used was InterProSurf (http://curie.utmb.edu/). This method is based on solvent-accessible surface area of residues in isolated proteins, a propensity scale for interface residues, and a clustering algorithm to identify surface regions with residues of high interface propensities [33]. Only the first five clusters were considered. All analysis was done using the default parameters for each algorithm. In the present work, a residue in the surface should be identified as at least by two of the above mentioned approaches to be considered an interaction site.

Evaluation of Interface Proximity

To evaluate whether the proximity of an aggregation-prone region to a given real interface is specific or the sequence stretch is as close to any other patch of the same size in the protein surface, we have defined the Interface Proximity Index: IPI

$$IPI = 1-(SP/IP)$$

$$IP = Interface\ Proximity = nR/nHS$$

$$SP = Surface\ Proximity = \sum_{nS=1}^{100} nS/nHS \Big/ 100$$

nR = number of residues in the aggregation-prone region at less than 3 Å from the interface.

nHS = number of residues in the aggregation-prone region.

nS = number of residues in the aggregation-prone region at less than 3 Å from a randomly chosen protein surface that does not include the interface.

Each random surface was generated by an aleatory selection of a number of solvent exposed residues equal to the number of residues constituting the real interface. One hundred random surfaces were generated for each aggregation-prone region analyzed.

Solvent-accessible and buried residues in the monomeric complex subunits where identified using the PISA server at the European Bioinformatics Institute (http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html).

An $IPI \leq 0$ indicates that the aggregation-prone region is equally or less close to the interface than to the rest of the surface. An $IPI > 0$ indicates that the aggregation-prone region is closer to the interface than to the rest of the surface, e. g., an $IPI = 0.5$ indicates that the aggregation-prone region is half as far from the interface than from the rest of the surface. The maximum value for IPI is 1.

Figures were were generated with the Swiss-PDB viewer program (http://spdbv.vital-it.ch) and rendered with POV (Persistance of Vision).

Acknowledgments

We thank Daniel Fernandez for the help with the ODA analysis and with the first draft of the present manuscript.

Author Contributions

Conceived and designed the experiments: SV. Performed the experiments: VC SV. Analyzed the data: VC SV. Wrote the paper: SV.

References

1. Chiti F, Dobson CM (2006) Protein misfolding, functional amyloid, and human disease. AnnuRevBiochem 75: 333.
2. Bellotti V, Chiti F (2008) Amyloidogenesis in its biological environment: challenging a fundamental issue in protein misfolding diseases. Curr Opin Struct Biol 18: 771–779.
3. Fernandez-Busquets X, de Groot NS, Fernandez D, Ventura S (2008) Recent structural and computational insights into conformational diseases. Curr Med Chem 15: 1336–1349.
4. Soto C, Estrada LD (2008) Protein misfolding and neurodegeneration. Arch Neurol 65: 184–189.
5. Pepys MB (2006) Amyloidosis. Annu Rev Med 57: 223–241.
6. Selkoe DJ (2003) Folding proteins in fatal ways. Nature 426: 900–904.
7. Ventura S (2005) Sequence determinants of protein aggregation: tools to increase protein solubility. Microbial Cell Factories 4: 11.
8. Sasahara K, Yagi H, Naiki H, Goto Y (2007) Heat-induced conversion of beta(2)-microglobulin and hen egg-white lysozyme into amyloid fibrils. J Mol Biol 372: 981–991.
9. Dumoulin M, Canet D, Last AM, Pardon E, Archer DB, et al. (2005) Reduced global cooperativity is a common feature underlying the amyloidogenicity of pathogenic lysozyme mutations. J Mol Biol 346: 773–788.
10. Hurshman Babbes AR, Powers ET, Kelly JW (2008) Quantification of the thermodynamically linked quaternary and tertiary structural stabilities of transthyretin and its disease-associated variants: the relationship between stability and amyloidosis. Biochemistry 47: 6969–6984.
11. DiDonato M, Craig L, Huff ME, Thayer MM, Cardoso RM, et al. (2003) ALS mutants of human superoxide dismutase form fibrous aggregates via framework destabilization. J Mol Biol 332: 601–615.
12. 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.
13. Chiti F, Taddei N, Bucciantini M, White P, Ramponi G, et al. (2000) Mutational analysis of the propensity for amyloid formation by a globular protein. Embo J 19: 1441–1449.
14. Pallares I, Vendrell J, Aviles FX, Ventura S (2004) Amyloid fibril formation by a partially structured intermediate state of alpha-chymotrypsin. J Mol Biol 342: 321–331.
15. Fandrich M, Fletcher MA, Dobson CM (2001) Amyloid fibrils from muscle myoglobin. Nature 410: 165–166.
16. Dobson CM (2004) Principles of protein folding, misfolding and aggregation. Semin Cell Dev Biol 15: 3–16.
17. 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.
18. Ivanova MI, Sawaya MR, Gingery M, Attinger A, Eisenberg D (2004) An amyloid-forming segment of {beta}2-microglobulin suggests a molecular model for the fibril. PNAS %R 101073/pnas0403756101 101: 10584–10589.
19. 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.
20. Caflisch A (2006) Computational models for the prediction of polypeptide aggregation propensity. Curr Opin Chem Biol 10: 437–444.
21. Rousseau F, Schymkowitz J, Serrano L (2006) Protein aggregation and amyloidosis: confusion of the kinds? Curr Opin Struct Biol 16: 118–126.
22. Chiti F, Dobson CM (2009) Amyloid formation by globular proteins under native conditions. Nat Chem Biol 5: 15–22.
23. Kennedy D, Norman C (2005) What don't we know? Science 309: 75.
24. Bogan AA, Thorn KS (1998) Anatomy of hot spots in protein interfaces. J Mol Biol 280: 1–9.
25. Ma B, Wolfson HJ, Nussinov R (2001) Protein functional epitopes: hot spots, dynamics and combinatorial libraries. Curr Opin Struct Biol 11: 364–369.

26. Ma B, Elkayam T, Wolfson H, Nussinov R (2003) Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc Natl Acad Sci U S A 100: 5772–5777.
27. Keskin O, Nussinov R, Gursoy A (2008) Prism: protein-protein interaction prediction by structural matching. Methods Mol Biol 484: 505–521.
28. Ofran Y, Rost B (2007) Protein-protein interaction hotspots carved into sequences. PLoS Comput Biol 3: e119.
29. Hoskins J, Lovell S, Blundell TL (2006) An algorithm for predicting protein-protein interaction sites: Abnormally exposed amino acid residues and secondary structure elements. Protein Sci 15: 1017–1029.
30. Sikic M, Tomic S, Vlahovicek K (2009) Prediction of protein-protein interaction sites in sequences and 3D structures by random forests. PLoS Comput Biol 5: e1000278.
31. Murakami Y, Jones S (2006) SHARP2: protein-protein interaction predictions using patch analysis. Bioinformatics 22: 1794–1795.
32. Fernandez-Recio J, Totrov M, Skorodumov C, Abagyan R (2005) Optimal docking area: a new method for predicting protein-protein interaction sites. Proteins 58: 134–143.
33. Negi SS, Schein CH, Oezguen N, Power TD, Braun W (2007) InterProSurf: a web server for predicting interacting sites on protein surfaces. Bioinformatics 23: 3397–3399.
34. Tuncbag N, Kar G, Keskin O, Gursoy A, Nussinov R (2009) A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. Brief Bioinform.
35. Ma B, Nussinov R (2007) Trp/Met/Phe hot spots in protein-protein interactions: potential targets in drug design. Curr Top Med Chem 7: 999–1005.
36. Monsellier E, Ramazzotti M, Taddei N, Chiti F (2008) Aggregation propensity of the human proteome. PLoS Comput Biol 4: e1000199.
37. DuBay KF, Pawar AP, Chiti F, Zurdo J, Dobson CM, et al. (2004) Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains. J Mol Biol 341: 1317–1326.
38. 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.
39. 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.
40. 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.
41. Galzitskaya OV, Garbuzynskiy SO, Lobanov MY (2006) Prediction of amyloidogenic and disordered regions in protein chains. PLoS Comput Biol 2: e177.
42. 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.
43. Zhang Z, Chen H, Lai L (2007) Identification of amyloid fibril-forming segments based on structure and residue-based statistical potential. Bioinformatics 23: 2218–2225.
44. Trovato A, Seno F, Tosatto SC (2007) The PASTA server for protein aggregation prediction. Protein Eng Des Sel 20: 521–523.
45. Monsellier E, Chiti F (2007) Prevention of amyloid-like aggregation as a driving force of protein evolution. EMBO Rep 8: 737–742.
46. Koch KM (1992) Dialysis-related amyloidosis. Kidney Int 41: 1416–1429.
47. Kozhukh GV, Hagihara Y, Kawakami T, Hasegawa K, Naiki H, et al. (2002) Investigation of a peptide responsible for amyloid fibril formation of beta 2-microglobulin by achromobacter protease I. J Biol Chem 277: 1310–1315.
48. Jones S, Manning J, Kad NM, Radford SE (2003) Amyloid-forming peptides from beta2-microglobulin-Insights into the mechanism of fibril formation in vitro. J Mol Biol 325: 249–257.
49. Kourilsky P, Claverie JM (1989) MHC restriction, alloreactivity, and thymic education: a common link? Cell 56: 327–329.
50. Tysoe-Calnon VA, Grundy JE, Perkins SJ (1991) Molecular comparisons of the beta 2-microglobulin-binding site in class I major-histocompatibility-complex alpha-chains and proteins of related sequences. Biochem J 277(Pt 2): 359–369.
51. Khan AR, Baker BM, Ghosh P, Biddison WE, Wiley DC (2000) The structure and stability of an HLA-A*0201/octameric tax peptide complex with an empty conserved peptide-N-terminal binding site. J Immunol 164: 6398–6405.
52. Enns CA (2001) Pumping iron: the strange partnership of the hemochromatosis protein, a class I MHC homolog, with the transferrin receptor. Traffic 2: 167–174.
53. Pietrangelo A (2006) Hereditary hemochromatosis. Biochim Biophys Acta 1763: 700–710.
54. Lebron JA, Bennett MJ, Vaughn DE, Chirino AJ, Snow PM, et al. (1998) Crystal structure of the hemochromatosis protein HFE and characterization of its interaction with transferrin receptor. Cell 93: 111–123.
55. Jahn TR, Parker MJ, Homans SW, Radford SE (2006) Amyloid formation under physiological conditions proceeds via a native-like folding intermediate. Nat Struct Mol Biol 13: 195–201.
56. Eakin CM, Berman AJ, Miranker AD (2006) A native to amyloidogenic transition regulated by a backbone trigger. Nat Struct Mol Biol 13: 202–208.
57. Floege J, Ehlerding G (1996) Beta-2-microglobulin-associated amyloidosis. Nephron 72: 9–26.
58. Connors LH, Lim A, Prokaeva T, Roskens VA, Costello CE (2003) Tabulation of human transthyretin (TTR) variants, 2003. Amyloid 10: 160–184.
59. Jaroniec CP, MacPhee CE, Bajaj VS, McMahon MT, Dobson CM, et al. (2004) 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 101: 711–716.
60. Jarvis JA, Kirkpatrick A, Craik DJ (1994) 1H NMR analysis of fibril-forming peptide fragments of transthyretin. Int J Pept Protein Res 44: 388–339.
61. Hamilton JA, Steinrauf LK, Braden BC, Liepnieks J, Benson MD, et al. (1993) The x-ray crystal structure refinements of normal human transthyretin and the amyloidogenic Val-30→Met variant to 1.7-A resolution. J Biol Chem 268: 2416–2424.
62. Foss TR, Wiseman RL, Kelly JW (2005) The pathway by which the tetrameric protein transthyretin dissociates. Biochemistry 44: 15525–15533.
63. Hornberg A, Eneqvist T, Olofsson A, Lundgren E, Sauer-Eriksson AE (2000) A comparative analysis of 23 structures of the amyloidogenic protein transthyretin. J Mol Biol 22: 649–669.
64. Hammarstrom P, Schneider F, Kelly JW (2001) Trans-suppression of misfolding in an amyloid disease. Science 293: 2459–2462.
65. Deng HX, Hentati A, Tainer JA, Iqbal Z, Cayabyab A, et al. (1993) Amyotrophic lateral sclerosis and structural defects in Cu,Zn superoxide dismutase. Science 261: 1047–1051.
66. Elam JS, Taylor AB, Strange R, Antonyuk S, Doucette PA, et al. (2003) Amyloid-like filaments and water-filled nanotubes formed by SOD1 mutant proteins linked to familial ALS. Nat Struct Biol 10: 461–467.
67. Chabry J, Ratsimanohatra C, Sponne I, Elena PP, Vincent JP, et al. (2003) In vivo and in vitro neurotoxicity of the human prion protein (PrP) fragment P118–135 independently of PrP expression. J Neurosci 23: 462–469.
68. Ince PG, Shaw PJ, Slade JY, Jones C, Hudgson P (1996) Familial amyotrophic lateral sclerosis with a mutation in exon 4 of the Cu/Zn superoxide dismutase gene: pathological and immunocytochemical changes. Acta Neuropathol 92: 395–403.
69. Stathopulos PB, Rumfeldt JA, Scholz GA, Irani RA, Frey HE, et al. (2003) Cu/Zn superoxide dismutase mutants associated with amyotrophic lateral sclerosis show enhanced formation of aggregates in vitro. Proc Natl Acad Sci U S A 100: 7021–7026.
70. Hough MA, Grossmann JG, Antonyuk SV, Strange RW, Doucette PA, et al. (2004) Dimer destabilization in superoxide dismutase may result in disease-causing properties: structures of motor neuron disease mutants. Proc Natl Acad Sci U S A 101: 5976–5981.
71. Ray SS, Lansbury PT Jr (2004) A possible therapeutic target for Lou Gehrig's disease. Proc Natl Acad Sci U S A 101: 5701–5702.
72. Chattopadhyay M, Durazo A, Sohn SH, Strong CD, Gralla EB, et al. (2008) Initiation and elongation in fibrillation of ALS-linked superoxide dismutase. Proc Natl Acad Sci U S A 105: 18663–18668.
73. Sanchorawala V (2006) Light-chain (AL) amyloidosis: diagnosis and treatment. Clin J Am Soc Nephrol 1: 1331–1341.
74. Baden EM, Owen BA, Peterson FC, Volkman BF, Ramirez-Alvarado M, et al. (2008) Altered dimer interface decreases stability in an amyloidogenic protein. J Biol Chem 283: 15853–15860.
75. Eulitz M, Weiss DT, Solomon A (1990) Immunoglobulin heavy-chain-associated amyloidosis. Proc Natl Acad Sci U S A 87: 6542–6546.
76. Saphire EO, Parren PW, Pantophlet R, Zwick MB, Morris GM, et al. (2001) Crystal structure of a neutralizing human IGG against HIV-1: a template for vaccine design. Science 293: 1155–1159.
77. Pepys MB, Hawkins PN, Booth DR, Vigushin DM, Tennent GA, et al. (1993) Human lysozyme gene mutations cause hereditary systemic amyloidosis. Nature 362: 553–557.
78. Frare E, Mossuto MF, Polverino de Laureto P, Dumoulin M, Dobson CM, et al. (2006) Identification of the core structure of lysozyme amyloid fibrils by proteolysis. J Mol Biol 361: 551–561.
79. Tartaglia GG, Pawar AP, Campioni S, Dobson CM, Chiti F, et al. (2008) Prediction of aggregation-prone regions in structured proteins. J Mol Biol 380: 425–436.
80. Dumoulin M, Last AM, Desmyter A, Decanniere K, Canet D, et al. (2003) A camelid antibody fragment inhibits the formation of amyloid fibrils by human lysozyme. Nature 424: 783–788.
81. Paravastu AK, Leapman RD, Yau WM, Tycko R (2008) Molecular structural basis for polymorphism in Alzheimer's beta-amyloid fibrils. Proc Natl Acad Sci U S A 105: 18349–18354.
82. Hoyer W, Gronwall C, Jonsson A, Stahl S, Hard T (2008) Stabilization of a beta-hairpin in monomeric Alzheimer's amyloid-beta peptide inhibits amyloid formation. Proc Natl Acad Sci U S A 105: 5099–5104.
83. Fandrich M, Forge V, Buder K, Kittler M, Dobson CM, et al. (2003) Myoglobin forms amyloid fibrils by association of unfolded polypeptide segments. Proc Natl Acad Sci U S A 100: 15463–15468.
84. Riggs P (2001) Expression and purification of maltose-binding protein fusions. Curr Protoc Mol Biol Chapter 16: Unit16 16.
85. Holmgren A (1995) Thioredoxin structure and mechanism: conformational changes on oxidation of the active-site sulfhydryls to a disulfide. Structure 3: 239–243.
86. Qin J, Clore GM, Kennedy WM, Huth JR, Gronenborn AM (1995) Solution structure of human thioredoxin in a mixed disulfide intermediate complex with

its target peptide from the transcription factor NF kappa B. Structure 3: 289–297.
87. Otzen DE, Kristensen O, Oliveberg M (2000) Designed protein tetramer zipped together with a hydrophobic Alzheimer homology: a structural clue to amyloid assembly. Proc Natl Acad Sci U S A 97: 9907–9912.
88. Rousseau F, Serrano L, Schymkowitz JW (2006) How evolutionary pressure against protein aggregation shaped chaperone specificity. J Mol Biol 355: 1037–1047.
89. Haas AL, Siepmann TJ (1997) Pathways of ubiquitin conjugation. FASEB J 11: 1257–1268.
90. Kang RS, Daniels CM, Francis SA, Shih SC, Salerno WJ, et al. (2003) Solution structure of a CUE-ubiquitin complex reveals a conserved mode of ubiquitin binding. Cell 113: 621–630.
91. Sloper-Mould KE, Jemc JC, Pickart CM, Hicke L (2001) Distinct functional surface regions on ubiquitin. J Biol Chem 276: 30483–30489.
92. Zhanhua C, Gan JG, Lei L, Sakharkar MK, Kangueane P (2005) Protein subunit interfaces: heterodimers versus homodimers. Bioinformation 1: 28–39.
93. Pechmann S, Levy ED, Tartaglia GG, Vendruscolo M (2009) Physicochemical principles that regulate the competition between functional and dysfunctional association of proteins. Proc Natl Acad Sci U S A 106: 10159–10164.
94. Nelson R, Sawaya MR, Balbirnie M, Madsen AO, Riekel C, et al. (2005) Structure of the cross-beta spine of amyloid-like fibrils. Nature 435: 773–778.

VIII - ANNEX

doi:10.1016/j.jmb.2008.03.020 J. Mol. Biol. (2008) 378, 1116–1131

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

The *in Vivo* and *in Vitro* Aggregation Properties of Globular Proteins Correlate With Their Conformational Stability: The SH3 Case

Alba Espargaró[1]†, Virginia Castillo[1]†, Natalia S. de Groot[1] and Salvador Ventura[1,2]*

[1]*Departament de Bioquímica i Biologia Molecular, Facultat de Biociències, Universitat Autònoma de Barcelona, E-08193 Bellaterra, Spain*

[2]*Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, E-08193 Bellaterra, Spain*

Received 16 October 2007; received in revised form 13 March 2008; accepted 13 March 2008
Available online 19 March 2008

Edited by J. Weissman

Protein misfolding and deposition underlie an increasing number of debilitating human disorders and constitute a problem of major concern in biotechnology. In the last years, *in vitro* studies have provided valuable insights into the physicochemical principles underlying protein aggregation. Nevertheless, information about the determinants of protein deposition within the cell is scarce and only a few systematic studies comparing *in vitro* and *in vivo* data have been reported. Here, we have used the SH3 domain of α-spectrin as a model globular protein in an attempt to understand the relationship between protein aggregation in the test-tube and in the more complex cellular environment. The investigation of the aggregation in *Escherichia coli* of this domain and a large set of mutants, together with the analysis of their sequential and conformational properties allowed us to evaluate the contribution of different polypeptidic factors to the cellular deposition of globular proteins. The data presented here suggest that the rules that govern *in vitro* protein aggregation are also valid in *in vivo* contexts. They also provide relevant insights into intracellular protein deposition in both conformational diseases and recombinant protein production.

Keywords: protein aggregation; SH3 domains; protein stability; protein folding; protein production

Introduction

The presence of insoluble protein deposits in human tissues correlates with the development of conformational disorders such as Alzheimer's disease and Parkinson's disease.[1–3] Also, the aggregation of globular proteins into insoluble polypeptide chains during protein production in bacteria is of major concern in biotechnology, since it reduces significantly the spectrum of relevant recombinant proteins available for structural genomics or commercial use.[4–6] For globular proteins, the formation of aggregates is in most cases an outcome of improper folding, resulting in the accumulation of unfolded or partially folded intermediates that are thought to be critical soluble precursors of fibrillar aggregates in amyloid diseases or inclusion bodies in protein production.[7,8] This effect has been characterized extensively *in vitro* by protein engineering using both disease-related and non-related polypeptide models.[9–14] Fewer studies have addressed the influence of polypeptide properties on their aggregation within the cell,[15–19] and only a few have compared systematically the *in vitro* and *in vivo* aggregation behaviour of proteins.[17,20–22] The *in vivo* aggregation of polypeptides does not necessarily have to correlate with their *in vitro* properties, because the protein quality machinery modulates the accumulation of aggregation-prone polypeptidic chains by facilitating their folding, masking hydro-

*Corresponding author. Departament de Bioquímica i Biologia Molecular, Facultat de Biociències, Universitat Autònoma de Barcelona, E-08193 Bellaterra, Spain. E-mail address: salvador.ventura@uab.es.

† A.E. and V.C. contributed equally to this work.

Abbreviations used: ANS, 1-anilinonaphtalene-8-sulfonate; FTIR, Fourier-transformed infrared; HypF-N, N terminus of hydrogenase maturation protein; PI3-kinase, phosphoinositide-3-kinase; SPC-SH3, SH3 domain of α-spectrin; Th-T, thioflavin-T; TEM, transmission electron microscopy.

phobic regions and targeting improperly folded proteins towards degradation pathways.[23] Our understanding of how protein aggregation proceeds within the cell will benefit from an integration of the well-established physicochemical principles underlying protein self-assembly *in vitro* with data from studies addressing protein deposition in living organisms, using consistent protein models.

Here, we use the SH3 domain of α-spectrin (SPC-SH3) as a model of a globular protein in an attempt to understand the determinants of protein aggregation in the cellular environment. SPC-SH3 is a 62 residue polypeptide that folds into an orthogonal β-sandwich.[24] The folding determinants of this protein have been characterized extensively.[25,26] In particular, the role of the hydrophobic core and the distal β-hairpin in the conformational stability of this domain has been studied. Both regions have been shown to be part of the folding nucleus of SH3 domains.[26,27] The hydrophobic core of SPC-SH3 comprises nine residues: Val9, Ala11, Val23, Met25, Leu31, Leu33, Val44, Val53 and Val58. The thermodynamic stability of 20 different *de novo* designed divergent variants at the core, containing from a single mutation up to the complete core substituted (Table 1), have been deduced from their equilibrium denaturation curves in the presence of urea at physiological pH.[28] The mutants span a range of stability of 6 kcal/mol (Table 2). Kinetic analysis of core variants with stability similar to that of the wild-type form demonstrated that, in general, they exhibit accelerated unfolding and refolding rates. This is compatible with the presence of conformational strain in the mutant native states, as deletion of a methyl group at the core leads to deceleration of the unfolding reaction and results in increased stability.[28] Wild-type SPC-SH3 displays conformational strain at the distal loop, caused by having the solvent-exposed Asn47 at position I of a type II β-turn in a high-energy region of the Ramachandran plot.[24] Mutation of the neighbouring, solvent-exposed, Asp48 to Gly (D48G) changes the type II β-turn to a type I β-turn, relaxing the strain, and stabilizes the protein by around 1.7 kcal/mol.[29] Accordingly, the Asn47 to Gly substitution (N47G) was also expected to stabilize the protein. Unexpectedly, the change in free energy was small (0.4 kcal/mol). X-ray analysis showed that in this case the conformation of the β-turn did not change and there was no relaxation of the structure.[30] Mutation of the solvent-exposed residues Glu7 (β1) and Lys60 (β5) to Tyr on the D48G background (D48G(2Y)) destabilizes the protein by ~0.7 kcal/mol relative to the wild type form (this work). Overall, the present collection of SPC-SH3 mutants constitutes a remarkable set of related polypeptides displaying differential thermodynamic and conformational properties.

Escherichia coli has emerged as a fast, simple, biologically and biotechnologically relevant experimental model to study the structural/sequential constraints underlying protein deposition *in vivo*.[31] We describe the cytosolic expression in *E. coli* of the SPC-SH3 variants described above and correlate the *in vivo* and *in vitro* aggregation behaviour of these globular species with their conformational stability, predicted aggregation propensities, and several other properties to provide a rational dissection of the con-

Table 1. Comparison of the sequence of the 23 mutants in the present study with the wild type SPC-SH3 domain

	Sequence position										
Protein	**9**	**11**	**23**	**25**	**31**	**33**	**44**	**47**	**48**	**53**	**58**
SPC-SH3 WT	**Val**	**Ala**	**Val**	**Met**	**Leu**	**Leu**	**Val**	**Asn**	**Asp**	**Val**	**Val**
B4 I25V	**Val**	Val	**Val**	Val	**Leu**	**Leu**	Ile	**Asn**	**Asp**	**Val**	Leu
B4 I44V	**Val**	Val	**Val**	Ile	**Leu**	**Leu**	**Val**	**Asn**	**Asp**	**Val**	Leu
B5 I25V	**Val**	Val	Leu	Val	**Leu**	**Leu**	Ile	**Asn**	**Asp**	**Val**	Leu
B5 I44V	**Val**	Val	Leu	Ile	**Leu**	**Leu**	**Val**	**Asn**	**Asp**	**Val**	Leu
Best	Ile	Val	Leu	Ile	**Leu**	**Leu**	Ile	**Asn**	**Asp**	Ile	Ile
Best2	Leu	Val	Leu	Ile	**Leu**	Ile	Ile	**Asn**	**Asp**	Ile	Ile
Best4	**Val**	Val	**Val**	Ile	**Leu**	**Leu**	Ile	**Asn**	**Asp**	**Val**	Leu
Best5	**Val**	Val	Leu	Ile	**Leu**	**Leu**	Ile	**Asn**	**Asp**	**Val**	Leu
Best7	Leu	Val	Leu	Ile	**Leu**	Ile	**Val**	**Asn**	**Asp**	Ile	Ile
Best9	Ala	Val	Leu	Ile	Ile	Ile	Leu	**Asn**	**Asp**	Phe	Leu
C8A	**Val**	Val	Leu	Ile	**Leu**	**Leu**	**Val**	**Asn**	**Asp**	Ile	Leu
C8A I25V	**Val**	Val	Leu	Val	**Leu**	**Leu**	**Val**	**Asn**	**Asp**	Ile	Leu
C8A I53V	**Val**	Val	Leu	Ile	**Leu**	**Leu**	**Val**	**Asn**	**Asp**	**Val**	Leu
C8B	**Val**	Val	Leu	Leu	**Leu**	**Leu**	Ala	**Asn**	**Asp**	Phe	Leu
D48G	**Val**	**Ala**	**Val**	**Met**	**Leu**	**Leu**	**Val**	**Asn**	Gly	**Val**	**Val**
D48G(2Y)	**Val**	**Ala**	**Val**	**Met**	**Leu**	**Leu**	**Val**	**Asn**	Gly	**Val**	**Val**
M25A	**Val**	**Ala**	**Val**	Ala	**Leu**	**Leu**	**Val**	**Asn**	**Asp**	**Val**	**Val**
Max-A	Ala	Val	Leu	Val	Ile	Ile	Ala	**Asn**	**Asp**	Phe	Trp
Max-F	Ile	Val	Leu	Ile	**Leu**	**Leu**	**Val**	**Asn**	**Asp**	Phe	Leu
Max-I	Ile	Ile	Ile	Ile	**Leu**	**Leu**	Ile	**Asn**	**Asp**	Ile	Ile
Max-L	Leu	Val	Leu	Leu	**Leu**	**Leu**	**Val**	**Asn**	**Asp**	Leu	Leu
Max-W	Ile	**Ala**	Leu	Val	**Leu**	**Leu**	Ile	**Asn**	**Asp**	**Val**	Trp
N47G	**Val**	**Ala**	**Val**	**Met**	**Leu**	**Leu**	**Val**	Gly	**Asp**	**Val**	**Val**

Natural residues are shown in bold. Residues 9, 11, 23, 25, 31, 33, 44, 53 and 58 form the hydrophobic core of the spectrin SH3 domain. Positions 47 and 48 are located at the distal β-hairpin. The D48G (2Y) mutant has additional Tyr mutations at positions 7 and 60.

Table 2. Comparison of *in vivo* aggregation, intrinsic properties and conformational stabilities of SH3 domains in the present study

	Solubility[a]	ΔG[b] (kcal/mol)	$\Delta\Delta G^{b}_{mut\text{-}wt}$ (kcal/mol)	AGGRESCAN[c]	TANGO[d]	AI[e]	Instability[f]	Half-life[g] (h)	GRAVY[h]
MAXA	–	<0.0	<−3.5	−15	581.6	77.1	27.63	>10	−0.741
MAXW	–	−0.5±0.30	−4.1	−13.4	577.1	88.06	36.45	>10	−0.726
C8B	–	0.3±0.05	−3.3	−13.6	573.4	88.06	27.02	>10	−0.689
Best9	–	0.5±0.03	−3.1	−11.8	728.6	89.68	25.65	>10	−0.661
MAXL	–	0.8±0.05	−2.8	−11.9	657.9	99.03	25.65	>10	−0.64
MAXF	–	1.0±0.03	−2.6	−9.9	669.1	92.74	28.76	>10	−0.634
M25A	–	1.0±0.03	−2.6	−15.5	580.9	84.84	26.56	>10	−0.685
Best7	–	1.1±0.09	−2.5	−9.1	726.8	99.03	25.19	>10	−0.595
Best2	–	1.2±0.02	−2.4	−8.7	726.5	100.65	32.28	>10	−0.59
Best	–	1.6±0.04	−2	−8.7	669	100.65	32.28	>10	−0.59
MAXI	–	2.1±0.02	−1.5	−7.6	668.8	102.26	38.49	>10	−0.574
D48G2Y	–	2.9±0.03	−0.7	−4.3	744	83.23	26.72	>10	−0.556
C8A	+	3.5±0.10	−0.1	−10.2	669.5	97.42	22.08	>10	−0.611
SH3 WT	+	3.6±0.10	0	−13.9	582.7	83.23	26.56	>10	−0.684
Best5	+	3.8±0.10	0.2	−10.2	669.3	97.42	32.74	>10	−0.611
N47G	+	4.0±0.02	0.4	−12.7	582.7	83.23	25.19	>10	−0.634
Best4	+	4.1±0.08	0.5	−9.8	670.5	95.81	31.37	>10	−0.605
B5–I25V[1]	++	4.6±0.04	1	−10.5	669.7	95.81	33.65	>10	−0.616
B4–I25V[1]	++	5.1±0.03	1.5	−10.2	670.53	94.19	32.28	>10	−0.61
D48G	++	5.3±0.07	1.7	−11.8	577.6	83.23	25.35	>10	−0.634
C8A –I25V	++	5.3±0.10	1.7	−10.5	669.9	95.81	22.99	>10	−0.616
B5–I44V[1]	++	5.6±0.10	2	−10.5	669.6	95.81	25.65	>10	−0.616
C8A–I53V	++	5.6±0.10	2	−10.5	669.6	95.81	25.65	>10	−0.616
B4–I44V[1]	++	5.8±0.10	2.2	−10.2	670.7	94.19	24.29	>10	−0.61

B5 and B4 account for Best5 and Best4, respectively.

[a] −, +,and ++, indicate that for a given variant, <30%, 30–70% and >70% of the SH3 domain is present in the soluble fraction, respectively.

[b] From references 26, 28, 29 and 30.

[c] Normalized aggregation propensity. Higher values indicate higher aggregation propensities. Calculated using http://bioinf.uab.es/aggrescan/.

[d] β-Aggregation. Higher values indicate higher aggregation propensities. Calculated using http://tango.crg.es/.

[e] Aliphatic index. Calculated using ProtParam at http://www.expasy.ch/cgi-bin/protparam.

[f] Instability index. All proteins are classified as stable. Calculated using ProtParam.

[g] Estimated half-life in *E. coli, in vivo*. Calculated using ProtParam.

[h] Grand average of hydropathicity. Calculated using ProtParam.

tribution of polypeptidic intrinsic factors to the cellular deposition of globular proteins.

Results and Discussion

SPC-SH3 and its mutants exhibit different *in vivo* aggregation propensity

We expressed wild-type SPC-SH3 and 23 different mutants in the cytoplasm of *E. coli* cells. The mutations map two critical regions for SH3 conformation, folding and stability: the hydrophobic core (20 mutants) and the distal β-hairpin (three mutants) (Table 1 and Fig. 1). All clones expressed SH3 domains at comparable levels, seen as a large ≈6.5 kDa band stained with Coomassie brilliant blue in the whole cell extract as monitored by SDS-PAGE (Supplementary Data Fig. S1). Cells were lysed, the proteins in the soluble and insoluble fractions were separated by centrifugation and the percentage of SH3 domain in the pellet and in the supernatant quantified by scanning SDS-PAGE gels of the respective cell fractions (Fig. 2).

The proteins could be classified into three sets according to the distribution of the recombinant polypeptide between the soluble and insoluble fractions. (i) Five variants, including the wild-type SPC-SH3, were present in similar amounts in both fractions, indicating that a portion of the SH3 domain was soluble within the cell while an equivalent quantity aggregated. (ii) Twelve mutants aggregate readily *in vivo* and were found exclusively or predominantly in the insoluble fraction. iii) Seven mutants were represented only or mainly in the fraction corresponding to the soluble cytosolic protein, showing that they remained mostly soluble even when over-expressed. Overall, the mutants displayed differential *in vivo* aggregation propensities, some of them aggregating more than the natural domain and some displaying similar or even increased solubility (Table 2 and Fig. 3).

Chiti and co-workers performed a related study using HypF-N as a protein model.[20] In that case, however, all variants were initially expressed as fusions at the C terminus of glutathione-*S*-transferase. It is well documented that the presence of a tag in the N-terminal position can significantly modify the co-translational folding and *in vivo* solubility of polypeptides.[32] Besides, despite the fact that the study provided highly relevant insights on the relationship between *in vitro* and *in vivo* aggregation, only six out of 18 mutants could be purified, due to their intrinsic *in vitro* insolubility. Thus, the conformation, stability,

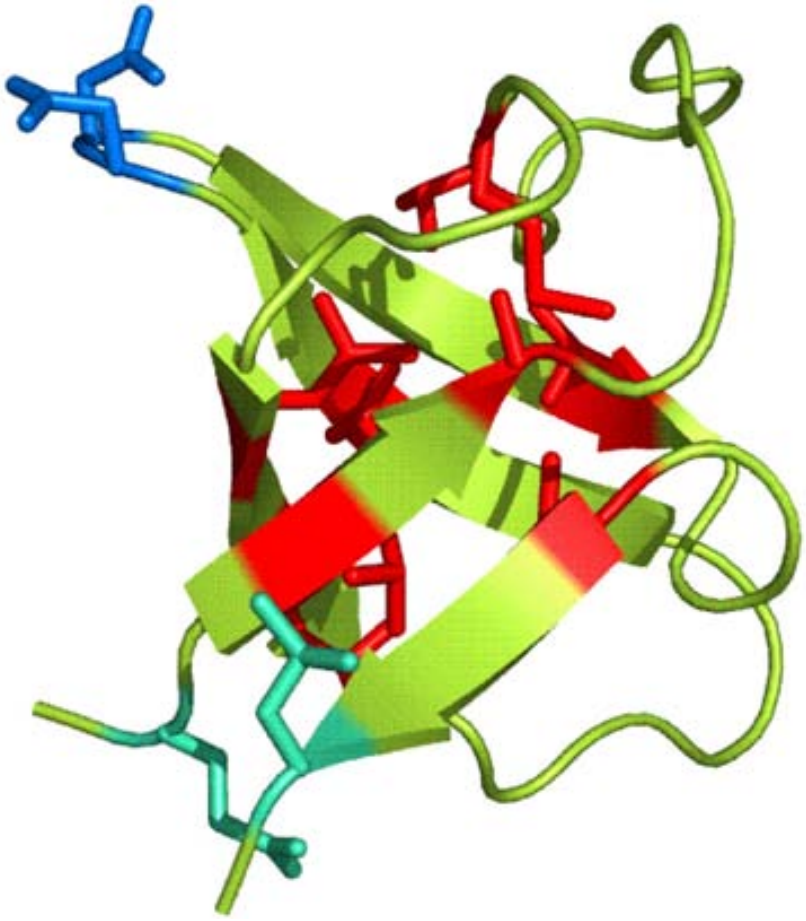


Fig. 1. A ribbon diagram of the α-spectrin SH3 domain. The side chains of residues mutated in the present study are represented. Residues at the hydrophobic core are shown in red, those in the distal β-hairpin are shown in blue and the rest are shown in magenta.

and *in vitro* characterization of aggregated states of most mutants could not be explored experimentally. Yet another difference with the present work is that no variant with significantly increased solubility relative to the wild-type form could be studied. Very recently, Fersht and co-workers reported an elegant study in which they could correlate the levels of folded recombinant p53 protein in *E. coli* with the *in vitro* thermodynamic stability of seven variants displaying both decreased and increased stability relative to the wild-type form. Again, the protein was expressed as a fusion, this time with enhanced green fluorescence protein. The foldability of p53 variants was evaluated by measuring the intracellular fluorescence of the fused reporter.[22] Nevertheless, one should be cautious when assigning intracellular fluorescence to properly folded and soluble proteins because, as we have shown using enhanced green fluorescence protein as a reporter, aggregated protein in bacterial inclusion bodies can be fluorescent,[33] and the levels of fluo-

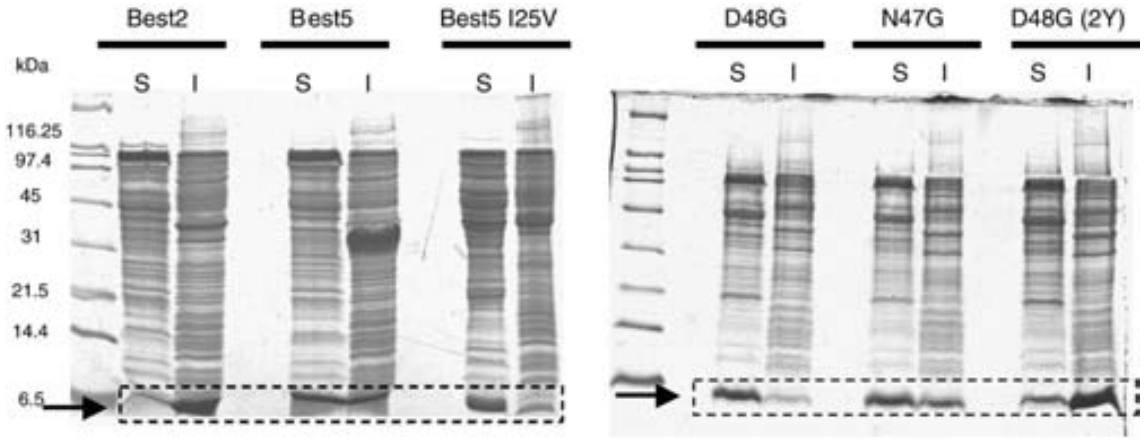


Fig. 2. SDS-PAGE analysis of cell fractions from representative SH3 mutants. Soluble (s) and insoluble (i) fractions of selected hydrophobic core (left) and distal β-hairpin (right) mutants are shown. The arrow indicates the bands stained with Coomassie brilliant blue that correspond to the SH3 domain molecular mass.

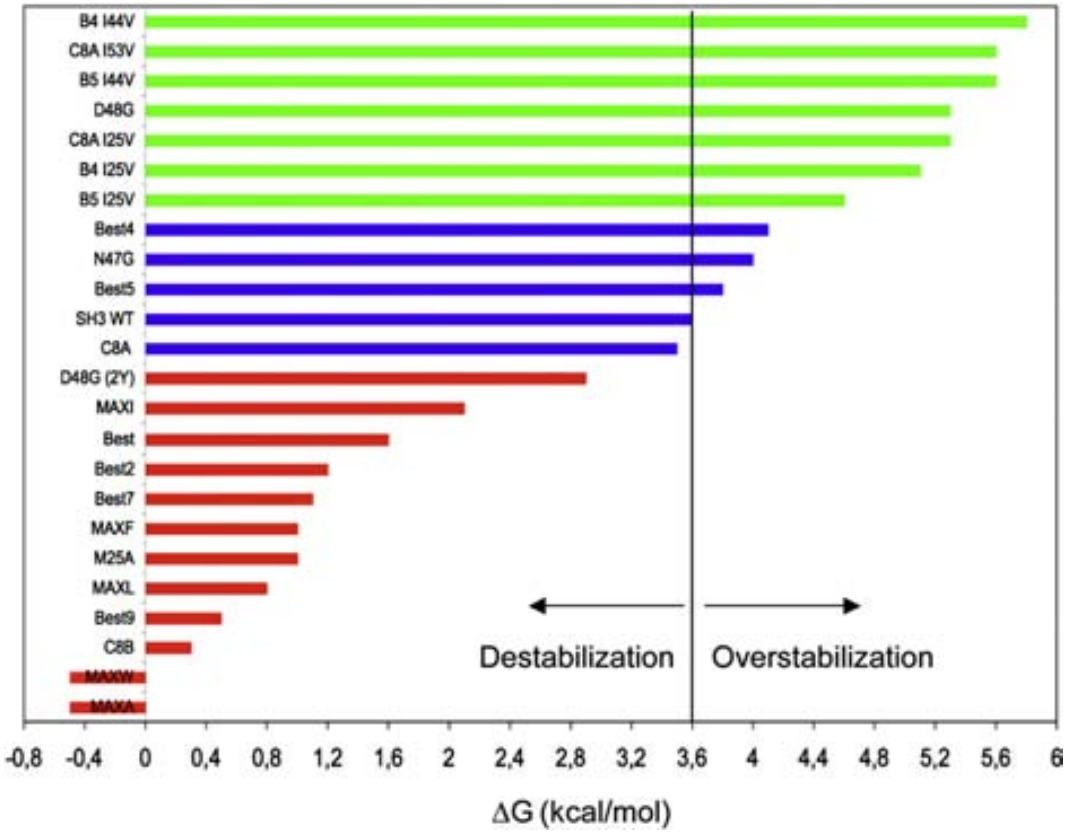


Fig. 3. Correlation between protein conformational stability and solubility. The bars indicate protein stability. Insoluble domains are shown in red and soluble domains are shown in green. Mutants present in similar amounts in the soluble and insoluble fractions are shown in blue. The line indicates wild-type stability (see the legend in Table 2).

rescence of the aggregated protein depend both on intrinsic factors, such as the sequence,[34] and extrinsic factors, such as the temperature.[35] In the present study, SH3 domains were expressed without any fusion, the quantity of soluble and aggregated protein was quantified directly and proteins with both increased and decreased stability were considered. As shown earlier, all the proteins could be purified either directly from the soluble fraction or upon denaturation of the insoluble fraction and *in vitro* refolding,[28–30] allowing the *in vitro* characterization of selected variants.

In vivo aggregation propensities of SPC-SH3 domains do not correlate with intrinsic sequential properties

The composition and arrangement of amino acids in the proteins has been suggested to be a major factor influencing their aggregation propensity. There have been many attempts to determine the relationship between the primary structure and the solubility status of a protein within the cell in order to enable rational identification of both natural and mutant protein candidates with a high level of solubility on over-expression and substitute common trial- and -error approaches.[36,37] Nevertheless, the expression conditions and the nature of proteins analysed in those studies were different, which is probably the reason for the dissimilarity between the results.[38] Thus, we reasoned that the present study, in which a large set of variants of a single domain were analysed in defined expression conditions, might help to decipher the intrinsic features influencing *in vivo* protein aggregation.

Among the physicochemical properties that can be read directly from the primary sequence, the aliphatic index has been described as a crucial determinant of solubility.[39] The aliphatic index corresponds to the relative volume occupied by aliphatic residues and has usually been associated with thermostability. Nevertheless, there is no correlation between this feature and the solubility of the proteins in our data set (Table 2). A higher level of predicted instability has been suggested to increase the propensity of proteins to be over-expressed in the soluble form, on the basis that a polypeptide with a shorter *in vivo* half-life has fewer long-lived intermediates.[39] However, all the SH3 domains were estimated to have similar half-lives in *E. coli* (>10 h) and no relationship between the instability index and the distribution of the different variants between the soluble and insoluble fractions was found (Table 2). Also, the enrichment in global hydrophobicity of the primary sequence has been suggested to be associated with an increased propensity to form aggregates inside *E. coli*;[36] yet such correlation does not apply for the present set of proteins, at least when hydrophobicity is measured as the grand average of hydropathicity of their sequences (Table 2). In addition, no relation-

ship between the volume of the hydrophobic core and SH3 solubility seems to exist.[28]

In vivo aggregation of globular SPC-SH3 domains does not correlate with predicted aggregation propensities of the unfolded state

Several studies indicate that there exist specific continuous protein segments that can nucleate the aggregation process when exposed to solvent in fully or partially unfolded contexts, suggesting a sequence-dependence of aggregation propensities.[40,41] Accordingly, in the last few years several computational approaches for detecting aggregation-prone regions and predicting relative polypeptide propensities to aggregate have been developed.[42,43]

We have shown that for the Alzheimer-related peptide Aβ-42, and a set of single-point mutants, *in vivo* aggregation rates upon expression inside bacterial cells correlated highly significantly with the aggregation propensities predicted theoretically from the sequence.[34] Thus, we tested whether computer-based approaches were also able to predict the differential *in vivo* aggregation of the SH3 domains in the present study. The sequences of the 24 proteins were analysed using AGGRESCAN[44] and TANGO[45] algorithms. No obvious correlation between the predicted global aggregations or the relative relevance of the different regions with predicted high aggregation propensity and the distribution of the protein forms between the cytosolic and insoluble fractions of the cell could be observed (Table 2).

The discrepancy between the accurate description of the in-the-cell aggregation propensities of Aβ-42 and the failure to predict the behaviour of SH3 domains by the above-mentioned, and related, algorithms can be rationalized according to the different conformational properties of both polypeptides. Aβ-42 is a mostly unstructured peptide in which aggregation-prone regions are already exposed to solvent and available for the establishment of inter-molecular contacts that may finally lead to the formation of aggregates. Thus, the effect of mutations that facilitate or hinder the conversion of the soluble unstructured state into oligomeric species have a direct impact on the average aggregation propensity of the peptide and can be explained in most cases by intrinsic factors. Because the same physicochemical principles appears to underlie the aggregation propensities of different polypeptides from unfolded states,[46] the effect of such mutations on aggregation propensity can be foreseen by most predictive algorithms.[42,43] These algorithms all assume that the aggregation-prone regions they detect are exposed to solvent at the time the aggregation begins and are thus in an at least partially unfolded context. In contrast to Aβ-42, SPC-SH3 is a small and compact globular protein in which aggregation-prone regions, if present, are likely to be blocked in the native state of the protein because their side chains are either hidden in the inner hydrophobic core or already involved in the network of contacts that stabilizes the protein. This explains why, once purified, SPC-SH3 is soluble *in vitro* in its native form at high concentrations (>10 mg/ml) for weeks in physiological conditions. Interestingly, the acid-unfolded state of the protein neither assembles into observable aggregates on the same time-scale and concentrations.[47] This suggests that, within the cell, SPC-SH3 domains do not aggregate by coalescence of fully folded or unfolded states and point to the formation and off-pathway association of partially folded intermediates during *in vivo* protein folding as the responsible for the differential intracellular aggregation properties of the SPC-SH3 domain and its mutants. This will explain why the *in vivo* aggregation propensities of SPC-SH3 variants cannot be anticipated directly from the linear information contained in the primary sequence either using simple approaches or employing sophisticated algorithms.

In vivo aggregation levels correlate with the conformational stability of globular SPC-SH3 domains

The conformational stability of polypeptides is expected to modulate the ensemble of partially folded conformations within the cell. Indeed, one should expect that destabilization of the, usually soluble, native state would likely increase the population of partially folded species, whereas over-stabilization of the native conformation would prevent the formation of significant levels of partially folded polypeptides in the cytosol. This implies that, if partially folded intermediates are the aggregation-prone species, protein stability will control the *in vivo* aggregation fate of globular proteins. To test if this applies to the SPC-SH3 domain, the distribution of the wild type and mutant proteins within the soluble and insoluble fractions was compared with their stability measured by urea denaturation at equilibrium in close to physiological conditions.[28] A striking correlation between solubility and conformational stability was found, independent of the place of mutation (Table 2 and Fig. 3). Interestingly, aggregation within the cell was not necessarily correlated to a large destabilization of the polypeptides resulting in only a minor population of the native state present at equilibrium. In this way, with the exception of MaxW and MaxA, all the mutants for which the majority of the protein was accumulated in the aggregated fraction display free energy changes of the unfolding transition in the absence of denaturant that are significantly different from zero. At the same time, wild-type protein stability, which in its natural context or *in vitro* under physiological conditions ensures solubility and functionality, rescues only about half of the protein from aggregation within the cell. This is in contrast to re-designed SH3 domains in which thermodynamic stability has been improved, since they remain almost completely in the soluble cytoplasmic fraction. According to our data, stabilization of the SH3 native state by more than 1 kcal/mol relative to the wild type (27%) ensures solubility whereas, in general, destabilization of more than 1 kcal/mol results in most of the protein being in the aggregated fraction

(Fig. 3). Lower levels of stabilization or destabilization result in a similar distribution of the protein between both fractions. Our data are in excellent agreement with that obtained by using the glutathione-*S*-transferase-HypF-N or enhanced green fluorescence protein-p53 models.[20,22] This confirms that in addition to the primary sequence, conformational stability is a major determinant for *in vivo* protein aggregation by modulating the population of partially folded intermediates with increased aggregation propensities.

SH3 *in vivo* deposition correlates with *in vitro* aggregation

SPC-SH3 and its mutants display a two-state transition unfolding curve when chemically denatured at low ionic strength.[28] Also, and in contrast to the homologous SH3 domain of PI3-kinase, SPC-SH3 does not aggregate into amyloid structures *in vitro* at highly acidic pH (2.0).[47] Nevertheless, it has been shown recently that destabilization of the SPC-SH3 folding nucleus by a point mutation (N47A) results in the population of partially folded conformers leading to the formation of amyloid aggregates at mild acidic pH (3.2).[48] Most of the residues mutated in the present study belong to the SH3 folding nucleus. Hence, we sought to explore whether there is any relationship between *in vitro* aggregation under mild denaturing conditions and the observed *in vivo* aggregation properties of SH3 domains.

We selected three representative mutants at the hydrophobic core corresponding to: (i) proteins accumulated mostly in the insoluble cell fraction and destabilized by more than 1 kcal/mol relative to the wild type (Best2); (ii) proteins distributed equally between the soluble and insoluble fractions with stability close to that of the wild-type protein (Best5); and (iii) proteins mostly soluble in *E. coli* and stabilized by more that 1 kcal/mol relative to the natural domain (Best5–I25V). The proteins were prepared at 1.7 mM (12 mg/ml) in buffer A (100 mM glycine buffer, pH 3.2, 100 mM NaCl)[48] and incubated at 25 °C. Under these conditions, the protein solution becomes slightly cloudy after dissolving Best2, whereas the solutions of Best5 and Best5–I25V remained clear. After incubation for one week, Best2 and Best5 solutions became gels. No macroscopic change was observed for Best5–I25V. Further quantification of the amount of aggregated protein by sample fractionation using sedimentation at 100,000*g* showed that most of the Best2 protein was aggregated (83%), about half of the Best5 protein remained soluble (53%) and a small amount of protein could be detected in the pelletable fraction of Best5–I25V sample (6%).

We studied the conformation of the incubated samples using far-UV CD (Fig. 4). Best2 and Best5 displayed CD spectra typical of β-sheet structure as revealed by the single negative band at 215–220 nm. These species were unable to revert to the native state on a reasonable time-scale after dilution of the sample, implying that a high-energy barrier separates these species under the conditions of the experiments. In contrast, the spectrum of the overstabilized Best5–I25V variant corresponds to a native SPC-SH3 conformation, with the characteristic maximum at 220 nm and minimum at 230 nm.

Addition of thioflavin-T (Th-T) to SH3 domains and incubation for one week in the conditions described above resulted in the following: Best2 and Best5 exhibited eightfold and fourfold increases in the emission at 482 nm, respectively. Only minor changes in the emission spectrum were detected when Th-T was added to the Best5-I25V domain (Fig. 4). We further investigated the properties of incubated SH3 domains by measuring their binding to Congo red, which has an absorbance maximum at 490 nm that shifts to red upon binding to amyloid-like material. In this case, only Best2 promoted the expected change in the absorbance spectra, suggesting an amyloid-like morphology of the aggregates (Supplementary Data Fig. S2). The same protein solutions were analysed by transmission electron microscopy (TEM) and the results are shown in Fig. 5. In agreement with the CD spectra and dye-binding properties, large amounts of fibrillar material were detected in solutions of Best2. Tight networks and isolated short, thin protofibrils of about 100 nm length and 5–10 nm widths were observed. This type of assembly remained without apparent change for weeks and no higher-order association into fibrils was observed in these conditions. The same kind of structures, although less abundant, could be observed for Best5. Most fields of Best5–I25V were devoid of particles, and the presence of aggregated fibrillar material was detected only occasionally.

To confirm that the observed correlation between *in vivo* and *in vitro* aggregation behaviour is not restricted to changes located at the hydrophobic core, we analysed the *in vitro* aggregation of N47G, D48G and D48G(2Y) variants, containing mutations at the distal β-hairpin, and the β1 and β5 strands. When incubated in buffer A, the formation of macroscopic aggregates was observed immediately for the destabilized D48G(2Y) variant. In contrast, the solutions of N47G, D48G mutants remained clear and without apparent precipitation during incubation for one week. Further quantification of the amount of aggregated protein showed that most of D48G(2Y) protein was aggregated (91%), and most of N47G (81%) and D48G (87%) SH3 domains remained soluble. Accordingly, the far-UV CD spectra of incubated N47G and D48G mutants correspond to native-like conformations (Fig. 4). The far-UV spectrum of D48G(2Y) corresponds to a mostly unfolded conformation but the CD signal was low, probably because much of the protein was precipitated from the solution. This effect could not be reversed by sonication for 10 min, indicating that the aggregates were highly stable. Fourier-transform infrared (FTIR) spectroscopy has proved to be a powerful tool for investigation of the structural characteristics of aggregated proteins. We used this technique to further analyse the structure of D48G(2Y) *in vitro* aggregates. SPC-SH3 is composed mainly of five β-sheets. According to this, the FTIR

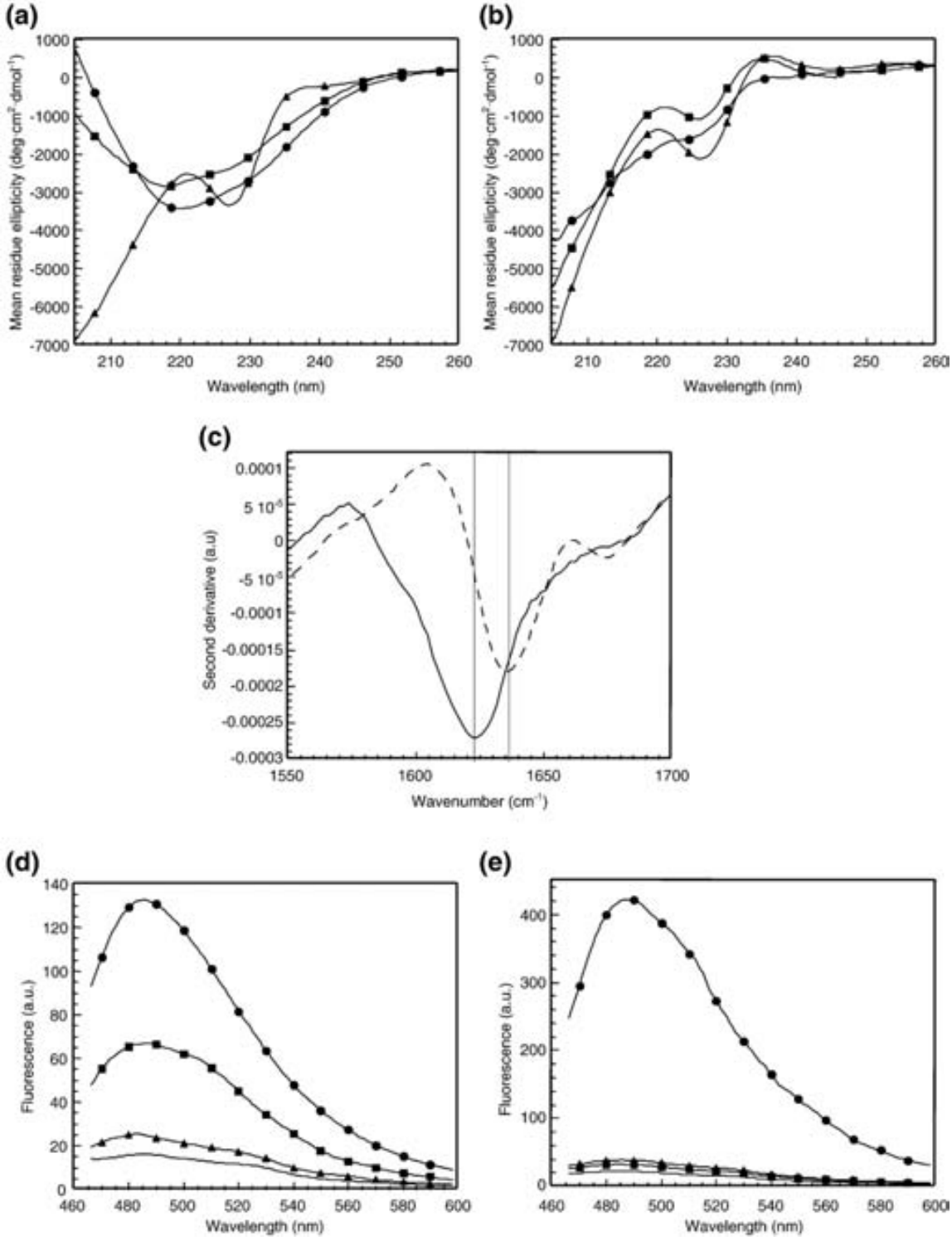


Fig. 4. Conformational properties of 1.7 mM (12 mg/ml) SH3 domains incubated at 25 °C for one week in 100 mM NaCl, 100 mM glycine (pH 3.2). (a and b) CD spectra in the far-UV region. (c) FTIR spectra in the amide I region. The positions of the spectral components were estimated from the second derivate analysis. (d and e) The fluorescence emission spectra of 40 μM thioflavin-T in the absence (no symbol) or in the presence of SH3 domains. The excitation wavelength was 440 nm. (a and d) Hydrophobic core SH3 mutants: Best2 (circles), Best5 (squares) and Best5–I25V (triangles). (b and e) Distal β-hairpin SH3 mutants: D48G(2Y) (circles), N47G (squares) and D48G (triangles). (c) Incubated D48G(2Y) (continuous lines) and native SPC-SH3 (broken line).

spectrum of native SPC-SH3 in the amide region I was dominated by a peak at 1636 cm^{-1} that arises from CO vibrations in intramolecular β-sheets (Fig. 4). The FTIR spectrum of D48G(2Y) aggregates exhibited a main peak at 1623 cm^{-1} (Fig. 4). This band is indicative of a predominance of extended intermo-

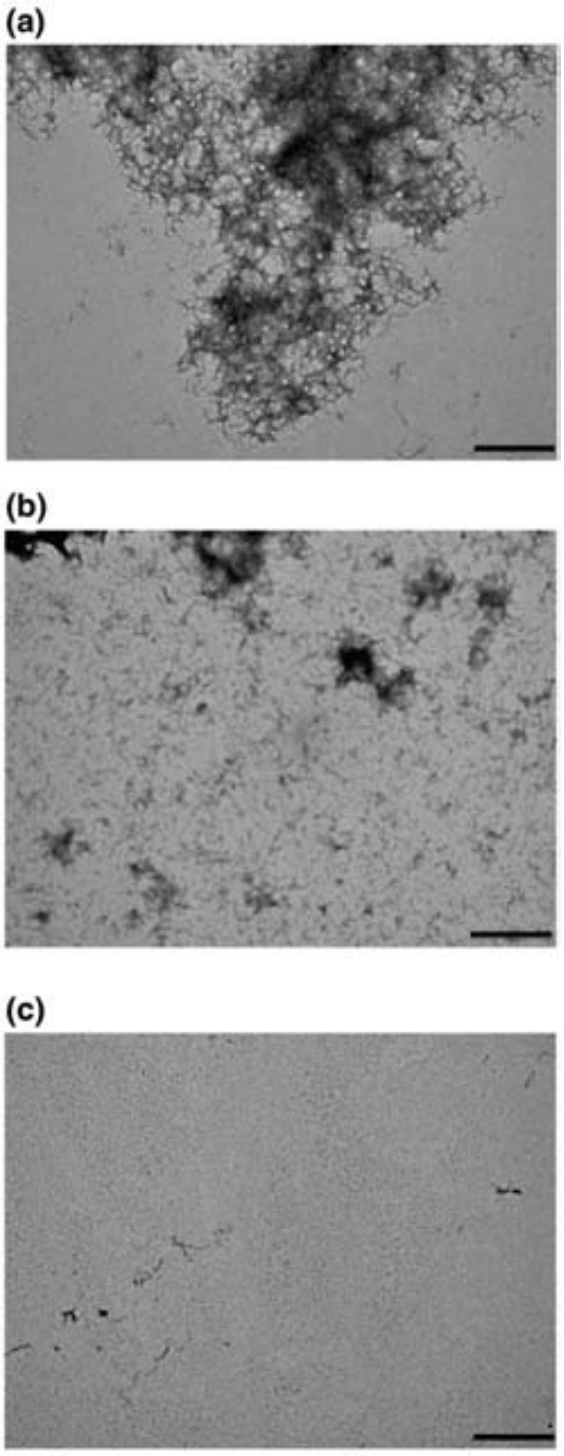


Fig. 5. Representative TEM images of hydrophobic core SH3 protein mutants incubated at 25 °C for one week. The samples were produced by tenfold dilutions of the original samples used for the incubation. (a) Best2; (b) Best5; and (c) Best5–I25V. The scale bar represents 500 nm.

lecular β-sheet structures in the aggregates. We further explored the structure of the incubated solutions by assaying their binding to Th-T and Congo red. No significant binding to such dyes was detected for N47G and D48G variants. In contrast, D48G (2Y) bind both Th-T and Congo red (Fig. 4 and Supplementary Data Fig. 2S). The affinity of D48G (2Y) for Th-T was surprisingly high, promoting a 20-fold increase of the fluorescence signal, suggesting an amyloid-like morphology of the aggregates. The same protein solutions were analysed by TEM, and the Results are shown in Fig. 6. Two different kinds of aggregates were detected in D48G(2Y) preparations: large amorphous aggregates that, in many cases, appear associated to protofibrillar material (Fig. 6a) and well-structured fibrils displaying a helical twisted morphology with ~60 nm width (Fig. 6b). This highly ordered fibrils have not been observed for any spectrin-SH3 domain and are likely responsible for the high Th-T affinity of this mutant. In agreement with the CD spectra and dye-binding properties, only very small aggregates were detected for the D48G mutant, whereas a reduced number of small fibrillar-like aggregates were observed in N47G samples (Fig. 6).

We also analysed *in vitro* aggregation of different SH3 domains at pH 5.0 and pH 7.0 (100 mM sodium phosphate buffer). Best5–I25V and D48G as well as Best2 and D48G(2Y) were selected as representative of stable and unstable mutants at the core and distal β-hairpin. They were incubated at a concentration of 1.7 mM (12 mg/ml) for one week at 25 °C. Macroscopic aggregates were observed immediately in Best2 and D48G(2Y) solutions at both pH values, whereas the over-stabilized Best5–I25V and D48G solutions remained clear. Upon incubation at pH 5.0 for one week, most of the Best2 (85%) and D48G(2Y) (81%) protein was aggregated. At pH 7.0, 67% of Best2 and 59% of D48G(2Y) proteins became insoluble, indicating that aggregation of unstable mutants occurs at close to physiological conditions. Best2 and D48G(2Y) pH 5.0 aggregates exhibited significant binding to Th-T (Supplementary Data Fig. S3), whereas the binding of pH 7.0 aggregates was much lower than that expected according to the amount of aggregated protein, suggesting that pH affects the amount of deposited protein and the conformational properties of the aggregates. No significant Th-T binding was observed for the over-stabilized Best5–I25V and D48G variants at any pH.

Overall, the *in vitro* properties of the analysed domains are in excellent agreement with their fractionation between soluble and insoluble compartments observed *in vivo*, indicating that, at least in certain cases, the data obtained using simplified non-cellular contexts can be relevant in *in vivo* environments.

In vivo formed SH3 aggregates display some amyloid-like properties

In order to test if *in vitro* and *in vivo* formed aggregates share morphological and/or conformational features, we purified and characterized the *in vivo* aggregates formed during the expression of the mostly insoluble Best2 and D48G(2Y) SH3 domains. Analysis by TEM shows that Best2 *in vivo* aggregates correspond to inclusion bodies (IBs) of around 1 μm size displaying a typical electro-dense ellipsoidal

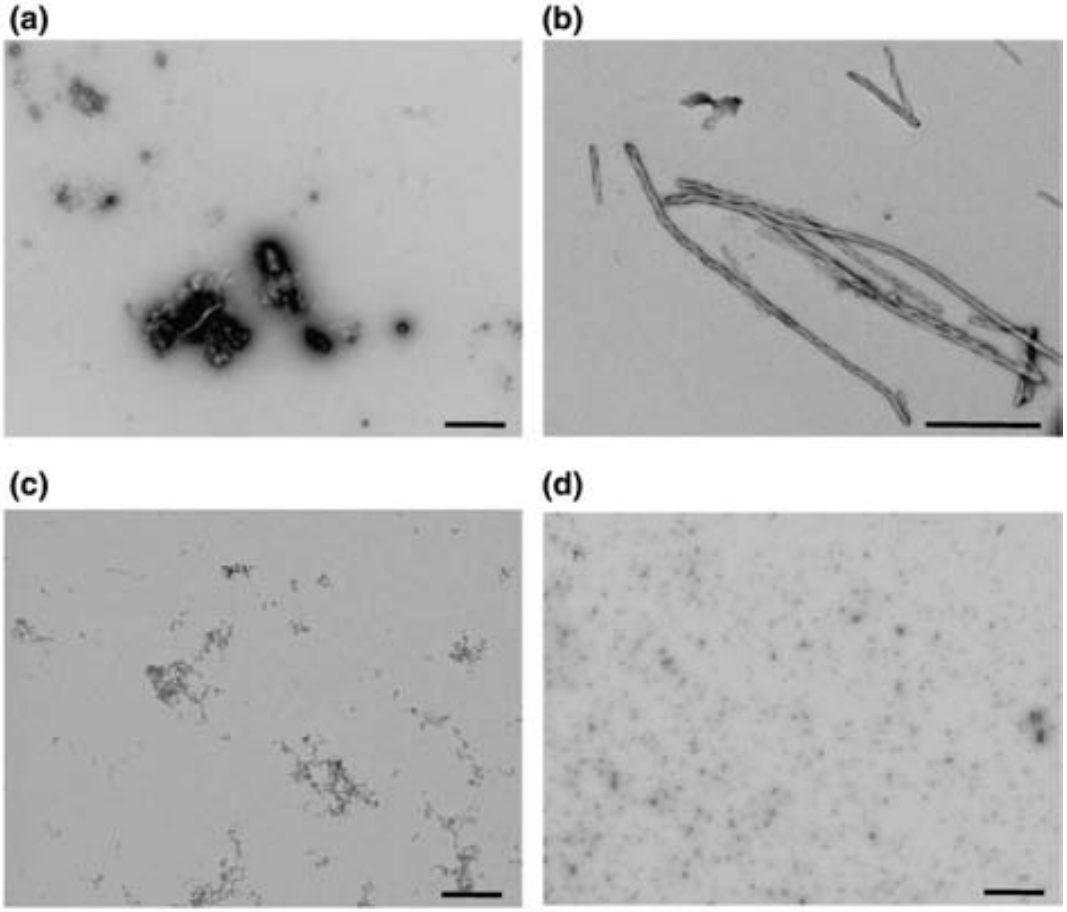


Fig. 6. Representative TEM images of distal β-hairpin SH3 mutants incubated at 25 °C for one week. The samples were produced by tenfold dilutions of the original samples used for incubation. (a and b) D48G(2Y); (c) N47G; and (d) D48G SH3 domains. The scale bar represents 1 μm.

morphology (Fig. 7a). The aggregates formed by D48G(2Y) display similar morphology, but short fibrillar material was usually detected in contact with them or in their vicinity (Fig. 7b). Bacterial IBs have often been regarded as disordered precipitates of non-specifically coagulated polypeptide chains. Nevertheless, we have shown that IBs might exhibit some degree of internal architecture resembling that of amyloid material.[31] We used FTIR to study the secondary structure of intracellular aggregates. The FTIR spectra of D48G(2Y) and Best2 aggregates exhibited major bands in the amide region I at 1626 cm^{-1} and 1628 cm^{-1}, respectively. This indicates the predominance of an extended intermolecular β-sheet architecture and suggests similar structural organization in the intracellular aggregates formed by different SH3 domains (Fig. 8). The displacement in the signal between aggregates might be related to differences in polypeptide packing or to the contribution to the spectrum of the fibrillar material detected in association with D48G(2Y) aggregates. We sought to test whether the detected β-sheet structure in SH3 *in vivo* aggregates might display amyloid-like features by measuring the binding to Th-T and Congo red. Addition of Th-T to purified intracellular SH3 aggregates resulted in the following: D48G(2Y) and Best2 exhibited sixfold and threefold increases in the emission at 482 nm, respectively (Fig. 8). Interestingly, the relative binding to Th-T exhibited by *in vivo* formed aggregates resemble closely that of the correspondent aggregates formed *in vitro*. Both aggregates also bind moderately to Congo red, as deduced from the shift to higher wavelengths of the maximum in the dye spectrum (Supplementary Data Fig. S4). Most amyloid fibrils are SDS-insoluble and resist high concentrations of urea or acid. To test whether the purified SH3 intracellular aggregates display such resistance, they were prepared at $A_{340\ nm}=1$ and incubated for 30 min in the presence of 5% (w/v) SDS, 9 M urea, a mixture of both reagents, or 70% (v/v) formic acid. In all these conditions, the turbidity of the solution disappeared completely at the end of the reaction, indicating the disintegration of the aggregates. SDS-PAGE analysis did not show the presence of SDS-resistant material in the aggregates. Thus, the stability of *in vivo* formed SH3 aggregates is clearly lower than that of typical fibrils. Still, the structural and dye-binding properties of SH3 IBs suggest that they might share common conformational features with amyloids.

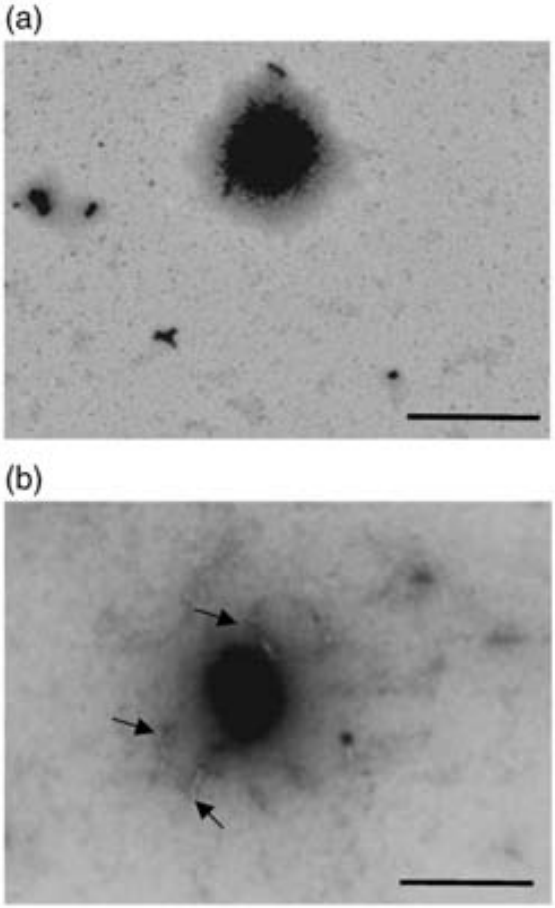


Fig. 7. Representative TEM images of intracellular SH3 aggregates. The samples were produced by tenfold dilutions of the original samples. (a) Best2 and (b) D48G(2Y) SH3 domains. The arrows indicate fibrillar material. The scale bar represents 1 μm.

In vitro aggregation correlates with the conformation and stability of globular SPC-SH3 domains

We sought to study the conformational features of proteins under *in vitro* conditions preceding the onset of aggregation, since these, or conformation-related, species might have a crucial role in the initial stages of *in vivo* protein aggregation by further evolving into aggregates. To this aim, fresh, 20 μM protein solutions of N47G, D48G D48G(2Y), Best2, Best5 and Best5–I25V mutants were prepared in buffer A and analysed by CD in the far-UV region. From the mutants at the hydrophobic core, Best2 displayed a non-native spectrum, whereas Best5 and Best5–I25V exhibited native-like conformations. All the mutants at the distal β-hairpin had spectra consistent with native-like conformations (Fig. 9).

Aggregation-prone intermediates usually expose hydrophobic patches to solvent, thus promoting protein self-assembly. The exposure of hydrophobic surfaces can be monitored by 1-anilinonaphtalene-8-sulfonate (ANS) binding. The incubation of 20 μM (0.14 mg/ml) destabilized Best2 and D48G(2Y) SH3 domains in the presence of 60 μM ANS promoted a pronounced increase in the fluorescence signal as well as a strong blue shift of the maximum wavelength (Fig. 9). Little binding of ANS to Best5, Best5–I25V, N47G and D48G was observed. Therefore, hydrophobic clusters are detected in the more unstable and aggregation-prone mutants. Interestingly, the presence of hydrophobic patches is detected both in mostly unfolded (Best2) and mostly folded contexts (D48G(2Y)) suggesting that the formation of SH3 intermediates with hydrophobic regions exposed to solvent might be a requirement for the transition toward β-sheet-rich aggregated states. Nevertheless, Best5 forms β-sheet-rich aggregates after prolonged incubation, despite the fact of displaying an initial native-like conformation at pH 3.2 and not exposing hydrophobic patches. Instead, Best5–I25V differing only in one methyl group at the hydrophobic core does not aggregate and keeps the native conformation at high concentrations after prolonged incubation. To test whether this discrepancy can be explained by their differential stability in the assay

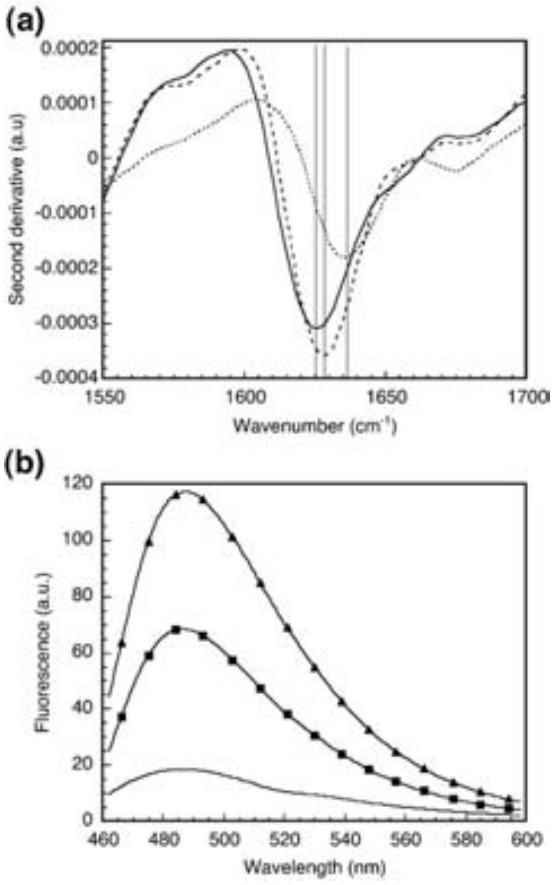


Fig. 8. Conformational properties of intracellular SH3 aggregates. (a) The FTIR spectrum of native SPC-SH3 (dotted line) and *in vivo* Best2 (broken line) and D48G(2Y) (continuous line) aggregates. The positions of the spectral components are estimated from the second derivate analysis. (b) Fluorescence emission spectra of thioflavin-T (40 μM) in the absence (no symbols) or in the presence of SH3 *in vivo* aggregates. The excitation wavelength was 440 nm; Best2 (squares) and D48G(2Y) (triangles).

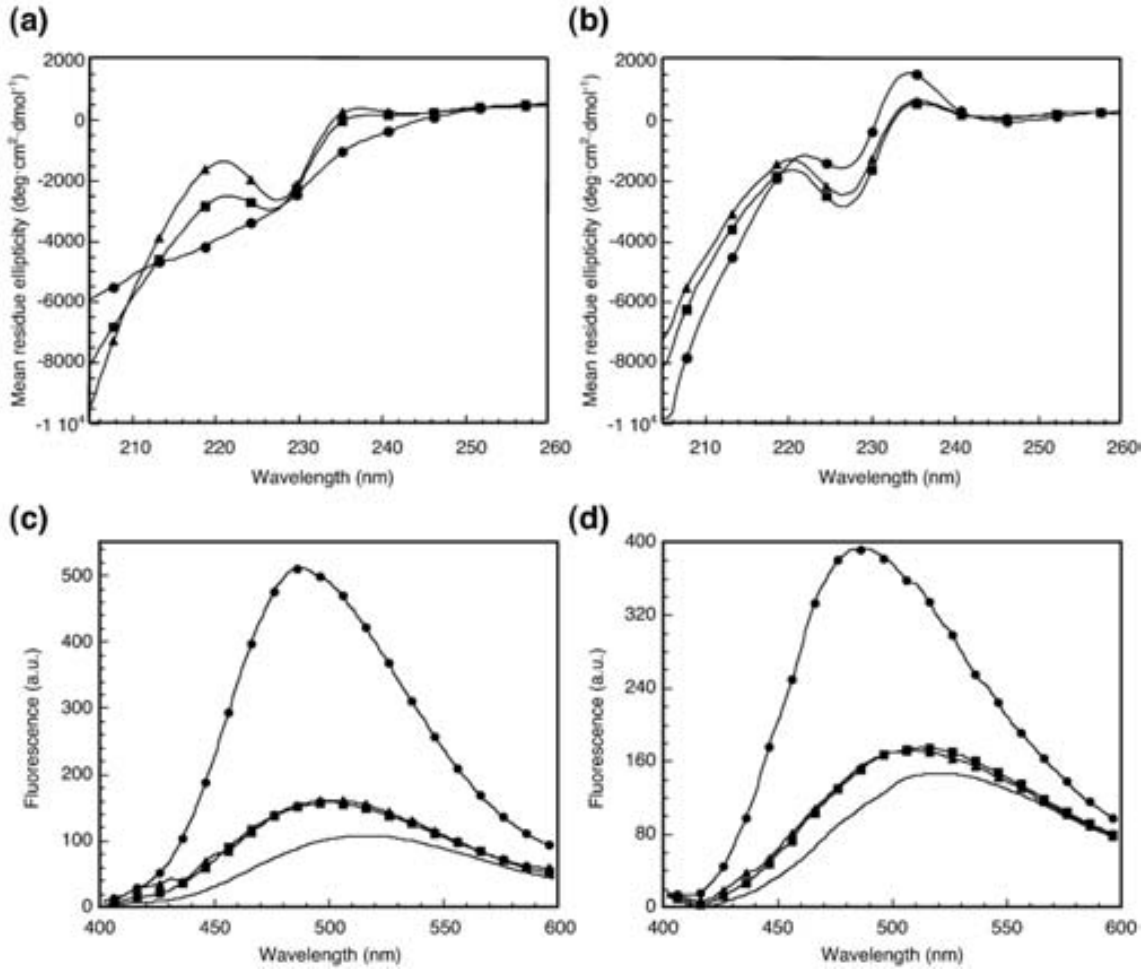


Fig. 9. Conformational properties of fresh SH3 domains solutions at 20 μM (0.14 mg/ml) and 25 °C in 100 mM NaCl, 100 mM glycine (pH 3.2). (a and b) CD spectra in the far-UV region. (c and d) Fluorescence emission spectra of 60 μM ANS in the absence (no symbol) or in the presence of SH3 domains. The excitation wavelength was 370 nm. (a and c) Hydrophobic core SH3 mutants; Best2 (circles), Best5 (squares) and Best5–I25V (triangles). (b and d) Distal β-hairpin SH3 mutants; D48G(2Y) (circles), N47G (squares) and D48G (triangles).

conditions, we analysed the thermal unfolding of these mutants at pH 3.2 in buffer A following the loss of the characteristic CD signal at 220 nm. Both mutants displayed a cooperative unfolding curve. Best5 was less stable ($T_m \approx 331$ K) than Best5–I25V ($T_m \approx 339$ K) in mild acidic conditions (Fig. 10), suggesting that the lower stability of its structure at pH 3.2, probably related to the presence of strain at the hydrophobic core, permits the transient sampling of aggregation-competent conformations and its subsequent self-assembly under conditions in which Best5–I25V, with an optimally packed interior, remains in a native-like conformation that protects it from deposition. Thus, as demonstrated within the cell, aggregation *in vitro* does not necessarily require a large destabilization that results in full unfolding of the protein at equilibrium. In globular proteins it appears that even subtle changes might determine the folding fate of the polypeptide, probably by modulating local fluctuations of the native backbone towards aggregation-prone conformations. Importantly, in other protein models, destabilization of the folding nucleus has been shown to lead to misfolding

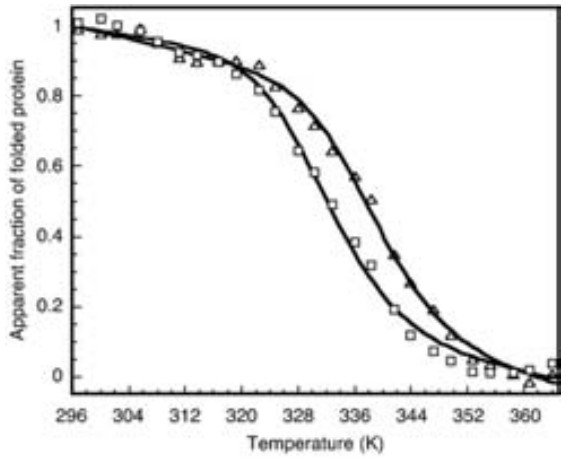


Fig. 10. Secondary structure changes on thermal denaturation of SH3 domains measured by CD. Ellipticity changes at 220 nm upon heating are shown for fresh solutions of 20 μM (0.14 mg/ml) Best5 (squares) and Best5–I25V (triangles) SH3 domains in 100 mM NaCl, 100 mM glycine (pH 3.2).

and subsequent in *vitro* aggregation, suggesting that the transient population of self-assembly-competent intermediate conformations might constitute a generic trigger of protein deposition.[49–52]

Conclusions

Overall, our data confirm that native state stability and cooperativity disfavours the aggregation of globular proteins, because it prevents local or global unfolding and thus the population of assembly-competent intermediates. This appears as a very successful evolutionary strategy to avoid aggregation, since few globular proteins aggregate from their stable native conformation in their natural environment.[6,53] Accordingly, in the conformational diseases in which globular proteins are implied, pathogenic mutations resulting in increased levels of aggregated protein usually destabilize the native folded conformation.[54,55] Importantly, as shown here, even highly conservative mutations can shift the equilibrium between folded and aggregated states. Our data support the use of approaches aimed at stabilizing the native state of globular amyloidogenic proteins as an efficient strategy for the therapeutics of amyloid diseases.[56]

From our study, it is likely that the same conformational constrains that determine *in vitro* aggregation might influence the *in vivo* fate of globular domains, thus validating the use of *in vitro* approaches to anticipate aggregation properties in the more complex cellular environment.

Aggregation reactions have serious ramifications in biotechnology, since, in many cases, they hinder the expression of sufficient quantities of a protein in the native, active state to allow for efficient production of protein-based drugs. The observed association of protein stability and solubility suggests that both properties can be targeted simultaneously during protein production. This is of significant interest for the biotechnology industry, since both features are usually a requirement for the biomedical application of polypeptides as drugs.

Materials and Methods

Protein expression and purification

Competent *E. coli* BL21 (DE3) cells were transformed with compatible pBAT4-derived plasmids encoding SPC-SH3 and its mutants. Transformed cells were incubated in Luria Bertani medium (with 100 μg/ml of ampicillin) overnight at 37 °C and then diluted to 1/100 (v/v). At $A_{600\ nm}=0.6$, the cultures were induced with 1 mM IPTG and incubated at 37 °C for 12 h. Soluble and insoluble fractions were isolated using the BugBuster® Protein Extraction Reagent from Novagene according to the procedure recommended by the manufacturer, and the proteins present in these fractions were resolved on Tricine-SDS/12% (w/v) PAGE gels. Gels were stained with Coomassie brilliant blue, scanned at high resolution and SH3 bands quantified with the Quantity One software from Bio Rad. SPC-SH3 and its mutants were purified as described.[26] Homogeneous samples of SH3 were dialysed extensively against pure water and lyophilized. Protein concentration was determined by measurement of absorbance at 280 nm using extinction coefficients as described.[57] All mutants were sequenced and protein identity was checked by mass spectrometry. Unless indicated otherwise, proteins were reconstituted in 100 mM glycine, 100 mM NaCl, pH 3.2 buffer for aggregation and conformational studies. Intracellular protein aggregates were purified from cell extracts by detergent-based procedures as described.[34] For the determination of inclusion body protein, these structures were resuspended in 6 M guanidinium hydrochloride for 1 h at 37 °C. The absorbance spectra of the different mutant solutions were recorded between 240 nm and 600 nm, and the scattering contribution to the signal was corrected. Protein concentration was calculated by using the molar absorption coefficient at 280 nm for the unfolded form of each SH3 mutant.

Binding to amyloid-diagnostic dyes

Samples were tested for Congo red binding by the spectroscopic band-shift assay as described.[58] Samples (5 μl) of 1.7 mM (12 mg/ml) protein solution were diluted with reaction buffer (5 mM sodium phosphate, 150 mM NaCl, pH 7.0) containing 15 μM Congo red. Samples were equilibrated for 5 min at 25 °C before analysis. Absorption spectra were collected together with negative control solutions of dye in the absence of protein and of protein samples in the absence of dye, subtracting the scattering contribution from the samples spectra, with a CARY-100 Varian spectrophotometer.

Th-T binding assays were carried out using samples of 5 μl from 1.7 mM (12 mg/ml) protein. These samples were diluted into buffer (10 mM sodium phosphate, 100 mM NaCl, pH 7.5) containing 40 μM Th-T, and adjusted to a final volume of 1 ml. Fluorescence data were collected after 5 min to ensure that thermal equilibrium had been achieved. Th-T was excited at 440 nm with a 2.5 nm slit width and the fluorescence emission was recorded at 485 nm with a 5 nm slit width, using a fluorescence spectrophotometer (Varian, Cary Eclipse). The same procedure was used to test the binding of Th-T and Congo red to *in vivo* formed aggregates.

Electron microscopy

Samples were incubated at 25 °C for one week before measurements for *in vitro* studies or purified from the insoluble fraction of cells for *in vivo* analysis. Samples were diluted tenfold with the same buffer and samples of 5–10 μl were placed on carbon-coated copper grids and left for 5 min. The grids were then washed and stained with 2% (w/v) uranyl acetate for another 5 min before analysis using a HITACHI H-7000 transmission electron microscope operating at an accelerating voltage of 75 kV.

Circular dichroism

CD experiments were performed with a Jasco J-715 spectropolarimeter. Measurements of the far-UV CD spectra (260–205 nm) were made with a 2 mm path-length quartz cuvette taking 10 μl samples from the aggregation mixture (proteins incubated at 25 °C during one week) and diluting them with the appropriate buffer (100 mM

glycine, 100 mM NaCl, pH 3.2) to a final volume of 500 μl (protein concentration of about 35 μM (0.25 mg/ml). Spectra were recorded at 25 °C. Non-incubated samples were analysed at 20 μM (0.14 mg/ml) final protein concentration in the same buffer. The resulting spectrum was the average of 20 scans. Thermal denaturation of SH3 domains was studied using a Jasco J-715 spectropolarimeter equipped with a thermostat controlled cell holder. Scans were conducted between 15 °C and 95 °C at a scan rate of 1 degC/min. Changes in CD signal at 220 nm were recorded. The protein concentration was 20 μM (0.14 mg/ml). 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 denaturation process.

FTIR spectroscopy analysis

SH3 samples were analysed using a Bruker Tensor 27 FT-IR spectrometer (Bruker Optics Inc) with a Golden Gate MKII ATR accessory. Each spectrum consists of 125 independent scans, measured at a spectral resolution of 2 cm^{-1} within the 1800–1500 cm−1 range. All spectral data were acquired and normalized using the OPUS MIR Tensor 27 software. Second derivatives of the spectra were used to determine the frequencies at which the different spectral components were located.

ANS binding assay

The fluorescence emission spectra of ANS in the presence and in the absence of protein were recorded with a fluorescence spectrophotometer (Varian, Cary Eclipse) at 25 °C. Proteins were diluted in 100 mM glycine, 100 mM NaCl, pH 3.2 buffer containing 60 μM ANS to obtain a final protein concentration of 20 μM (0.14 mg/ml). The ANS fluorescence was then read immediately using an excitation wavelength of 370 nm and the fluorescence emission was collected between 400 nm and 600 nm, with excitation and emission slit widths of 2.5 nm and 5 nm, respectively.

Acknowledgements

Most mutants were constructed originally in the laboratory of Dr Luis Serrano at EMBL-Heidelberg. We thank Prof. Aviles and Dr. Vendrell for lab facilities. This work was supported by grant BIO2007-68046 from the Ministerio de Ciencia y Tecnología, Spain, and by grant 2005SGR-00037 from AGAUR, Catalonia.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2008.03.020

References

1. Tan, S. Y. & Pepys, M. B. (1994). Amyloidosis. *Histopathology*, **25**, 403–414.
2. Sunde, M. & Blake, C. C. (1998). From the globular to the fibrous state: protein structure and structural conversion in amyloid formation. *Quart. Rev. Biophys.* **31**, 1–39.
3. Dobson, C. M. (2001). The structural basis of protein folding and its links with human disease. *Phil. Trans. Roy. Soc. B*, **356**, 133–145.
4. Fahnert, B., Lilie, H. & Neubauer, P. (2004). Inclusion bodies: formation and utilisation. *Adv. Biochem. Eng. Biotechnol.* **89**, 93–142.
5. Ventura, S. & Villaverde, A. (2006). Protein quality in bacterial inclusion bodies. *Trends Biotechnol.* **24**, 179–185.
6. Ventura, S. (2005). Sequence determinants of protein aggregation: tools to increase protein solubility. *Microbial Cell Factories*, **4**, 11.
7. Fink, A. L. (1998). Protein aggregation: folding aggregates, inclusion bodies and amyloid. *Fold. Des.* **3**, R9–R23.
8. Rinas, U., Hoffmann, F., Betiku, E., Estape, D. & Marten, S. (2007). 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.
9. Pallares, I., Vendrell, J., Aviles, F. X. & Ventura, S. (2004). Amyloid fibril formation by a partially structured intermediate state of α-chymotrypsin. *J. Mol. Biol.* **342**, 321–331.
10. Guijarro, J. I., Sunde, M., Jones, J. A., Campbell, I. D. & Dobson, C. M. (1998). Amyloid fibril formation by an SH3 domain. *Proc. Natl Acad. Sci. USA*, **95**, 4224–4228.
11. Ivanova, M. I., Thompson, M. J. & Eisenberg, D. (2006). A systematic screen of β(2)-microglobulin and insulin for amyloid-like segments. *Proc. Natl Acad. Sci. USA*, **103**, 4079–4082.
12. Hammarstrom, P., Sekijima, Y., White, J. T., Wiseman, R. L., Lim, A., Costello, C. E. *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.
13. McParland, V. J., Kad, N. M., Kalverda, A. P., Brown, A., Kirwin-Jones, P., Hunter, M. G. *et al.* (2000). Partially unfolded states of β-2-microglobulin and amyloid formation *in vitro*. *Biochemistry*, **39**, 8735–8746.
14. Zerovnik, E., Skarabot, M., Škerget, K., Giannini, S., Stoka, V., Jenko-Kokalj, S. & Staniforth, R. A. (2007). Amyloid fibril formation by human stefin B: influence of pH and TFE on fibril growth and morphology. *Amyloid*, **14**, 237–247.
15. Ghaemmaghami, S. & Oas, T. G. (2001). Quantitative protein stability measurement *in vivo*. *Nature Struct. Biol.* **8**, 879–882.
16. de Groot, N. S., Aviles, F. X., Vendrell, J. & Ventura, S. (2006). Mutagenesis of the central hydrophobic cluster in Aβ42 Alzheimer's peptide. Side-chain properties correlate with aggregation propensities. *FEBS J.* **273**, 658–668.
17. Wurth, C., Guimard, N. K. & Hecht, M. H. (2002). Mutations that reduce aggregation of the Alzheimer's Aβ42 peptide: an unbiased search for the sequence determinants of Aβ amyloidogenesis. *J. Mol. Biol.* **319**, 1279–1290.
18. Speed, M. A., Morshead, T., Wang, D. I. & King, J. (1997). Conformation of P22 tailspike folding and aggregation intermediates probed by monoclonal antibodies. *Protein Sci.* **6**, 99–108.
19. Speed, M. A., Wang, D. I. & King, J. (1996). Specific aggregation of partially folded polypeptide chains: the

molecular basis of inclusion body composition. *Nature Biotechnol.* **14**, 1283–1287.
20. Calloni, G., Zoffoli, S., Stefani, M., Dobson, C. M. & Chiti, F. (2005). Investigating the effects of mutations on protein aggregation in the cell. *J. Biol. Chem.* **280**, 10607–10613.
21. Ignatova, Z. & Gierasch, L. M. (2005). Aggregation of a slow-folding mutant of a β-clam protein proceeds through a monomeric nucleus. *Biochemistry*, **44**, 7266–7274.
22. Mayer, S., Rudiger, S., Ang, H. C., Joerger, A. C. & Fersht, A. R. (2007). Correlation of levels of folded recombinant p53 in *escherichia coli* with thermodynamic stability *in vitro*. *J. Mol. Biol.* **372**, 268–276.
23. Ma, Y. & Hendershot, L. M. (2001). The unfolding tale of the unfolded protein response. *Cell*, **107**, 827–830.
24. Musacchio, A., Noble, M., Pauptit, R., Wierenga, R. & Saraste, M. (1992). Crystal structure of a Src-homology 3 (SH3) domain. *Nature*, **359**, 851–855.
25. Martinez, J. C., Viguera, A. R., Berisio, R., Wilmanns, M., Mateo, P. L., Filimonov, V. V. & Serrano, L. (1999). Thermodynamic analysis of α-spectrin SH3 and two of its circular permutants with different loop lengths: discerning the reasons for rapid folding in proteins. *Biochemistry*, **38**, 549–559.
26. Martinez, J. C. & Serrano, L. (1999). The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nature Struct. Biol.* **6**, 1010–1016.
27. Grantcharova, V. P., Riddle, D. S., Santiago, J. V. & Baker, D. (1998). Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. *Nature Struct. Biol.* **5**, 714–720.
28. Ventura, S., Vega, M. C., Lacroix, E., Angrand, I., Spagnolo, L. & Serrano, L. (2002). Conformational strain in the hydrophobic core and its implications for protein folding and design. *Nature Struct. Biol.* **9**, 485–493.
29. Martinez, J. C., Pisabarro, M. T. & Serrano, L. (1998). Obligatory steps in protein folding and the conformational diversity of the transition state. *Nature Struct. Biol.* **5**, 721–729.
30. Vega, M. C., Martinez, J. C. & Serrano, L. (2000). Thermodynamic and structural characterization of Asn and Ala residues in the disallowed II′ region of the Ramachandran plot. *Protein Sci.* **9**, 2322–2328.
31. 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.
32. Davis, G. D., Elisee, C., Newham, D. M. & Harrison, R. G. (1999). New fusion protein systems designed to give soluble expression in *Escherichia coli*. *Biotechnol. Bioeng.* **65**, 382–388.
33. Garcia-Fruitos, E., Gonzalez-Montalban, N., Morell, M., Vera, A., Ferraz, R., Aris, A. *et al.* (2005). Aggregation as bacterial inclusion bodies does not imply inactivation of enzymes and fluorescent proteins. *Microb. Cell Fact.* **4**, 27.
34. de Groot, N. S. & Ventura, S. (2006). Protein activity in bacterial inclusion bodies correlates with predicted aggregation rates. *J. Biotechnol.* **125**, 110–113.
35. de Groot, N. S. & Ventura, S. (2006). Effect of temperature on protein quality in bacterial inclusion bodies. *FEBS Lett.* **580**, 6471–6476.
36. Luan, C. H., Qiu, S., Finley, J. B., Carson, M., Gray, R. J., Huang, W. *et al.* (2004). High-throughput expression of *C. elegans* proteins. *Genome Res.* **14**, 2102–2110.
37. Goh, C. S., Lan, N., Douglas, S. M., Wu, B., Echols, N., Smith, A. *et al.* (2004). Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis. *J. Mol. Biol.* **336**, 115–130.
38. Idicula-Thomas, S., Kulkarni, A. J., Kulkarni, B. D., Jayaraman, V. K. & Balaji, P. V. (2006). 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.
39. Idicula-Thomas, S. & 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.
40. Ivanova, M. I., Sawaya, M. R., Gingery, M., Attinger, A. & Eisenberg, D. (2004). An amyloid-forming segment of β2-microglobulin suggests a molecular model for the fibril. *Proc. Natl Acad. Sci. USA*, **101**, 10584–10589.
41. Ventura, S., Zurdo, J., Narayanan, S., Parreno, M., Mangues, R., Reif, B. *et al.* (2004). Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case. *Proc. Natl Acad. Sci. USA*, **101**, 7258–7263.
42. Bemporad, F., Calloni, G., Campioni, S., Plakoutsi, G., Taddei, N. & Chiti, F. (2006). Sequence and structural determinants of amyloid fibril formation. *Accts Chem. Res.* **39**, 620–627.
43. Caflisch, A. (2006). Computational models for the prediction of polypeptide aggregation propensity. *Curr. Opin. Chem. Biol.* **10**, 437–444.
44. Conchillo-Sole, O., de Groot, N. S., Aviles, F. X., Vendrell, J., Daura, X. & Ventura, S. (2007). AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics*, **8**, 65.
45. Fernandez-Escamilla, A. M., Rousseau, F., Schymkowitz, J. & Serrano, L. (2004). Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature Biotechnol.* **22**, 1302–1306.
46. 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.
47. Ventura, S., Lacroix, E. & Serrano, L. (2002). Insights into the origin of the tendency of the PI3-SH3 domain to form amyloid fibrils. *J. Mol. Biol.* **322**, 1147–1158.
48. Morel, B., Casares, S. & Conejero-Lara, F. (2006). A single mutation induces amyloid aggregation in the α-spectrin SH3 domain: analysis of the early stages of fibril formation. *J. Mol. Biol.* **356**, 453–468.
49. Dumoulin, M., Canet, D., Last, A. M., Pardon, E., Archer, D. B., Muyldermans, S. *et al.* (2005). Reduced global cooperativity is a common feature underlying the amyloidogenicity of pathogenic lysozyme mutations. *J. Mol. Biol.* **346**, 773–788.
50. Kim, Y., Wall, J. S., Meyer, J., Murphy, C., Randolph, T. W., Manning, M. C. *et al.* (2000). Thermodynamic modulation of light chain amyloid fibril formation. *J. Biol. Chem.* **275**, 1570–1574.
51. Ramirez-Alvarado, M., Merkel, J. S. & Regan, L. (2000). A systematic exploration of the influence of the protein stability on amyloid fibril formation *in vitro*. *Proc. Natl Acad. Sci. USA*, **97**, 8979–8984.
52. Smith, D. P., Jones, S., Serpell, L. C., Sunde, M. & Radford, S. E. (2003). A systematic investigation into the effect of protein destabilisation on beta 2-microglobulin amyloid formation. *J. Mol. Biol.* **330**, 943–954.

53. Selkoe, D. J. (2003). Folding proteins in fatal ways. *Nature*, **426**, 900–904.
54. Booth, D. R., Sunde, M., Bellotti, V., Robinson, C. V., Hutchinson, W. L., Fraser, P. E. *et al.* (1997). Instability, unfolding and aggregation of human lysozyme variants underlying amyloid fibrillogenesis. *Nature*, **385**, 787–793.
55. Takeuchi, M., Mizuguchi, M., Kouno, T., Shinohara, Y., Aizawa, T., Demura, M. *et al.* (2007). Destabilization of transthyretin by pathogenic mutations in the DE loop. *Proteins: Struct. Funct. Genet.* **66**, 716–725.
56. Cohen, F. E. & Kelly, J. W. (2003). Therapeutic approaches to protein-misfolding diseases. *Nature*, **426**, 905–909.
57. Gill, S. C. & Von Hippel, P. H. (1989). Calculation of protein extinction coefficients from amino acid sequence data. *Anal. Biochem.* **182**, 319–326.
58. Klunk, W. E., Pettegrew, J. W. & Abraham, D. J. (1989). Quantitative evaluation of congo red binding to amyloid-like proteins with a beta-pleated sheet conformation. *J. Histochem. Cytochem.* **37**, 1273–1281.

The N-terminal helix controls the transition between the soluble and amyloid states of an FF domain

Virginia Castillo[1,2], Fabrizio Chiti[3]* and Salvador Ventura[1,2]*

[1]Institut de Biotecnologia i de Biomedicina and

[2]Departament de Bioquímica i Biologia Molecular, Facultat de Biociències. Universitat Autònoma de Barcelona. E-08193 Bellaterra (Spain).

[3]Dipartimento di Scienze Biochimiche, Università di Firenze. 50134 Firenze (Italy).

* To whom correspondence should be addressed:

e-mail: fabrizio.chiti@unifi.it; Phone: +39 055 4598319; Fax +39 055 4598905

e-mail: salvador.ventura@uab.es; Phone: +34 93 5868956; Fax: +34 93 5811264

ABSTRACT

Protein misfolding and aggregation are interconnected processes linked to the onset of an increasing number of human nonneuropathic, systemic, and neurodegenerative disorders. In particular, misfolding of native α-helical structures and their self-assembly into nonnative intermolecular β-sheets has been proposed to trigger amyloid fibril formation in Alzheimer's and Parkinson's disorders. Here, we use a battery of biophysical techniques to elucidate the conformational conversion of native α-helices into amyloid fibrils using the all-α FF domain of yeast URN1 as a model system. FF domains are small four-helix bundle protein-protein interaction modules whose folding reaction has received much attention in recent times. However, their aggregation properties remained uncharacterized. We show that under mild denaturing conditions at low pH this FF domain self-assembles into amyloid fibrils. Theoretical and experimental dissection of the secondary structure elements in this domain indicates that the helix 1 at the N-terminus has both the highest α-helical and amyloid propensities, controlling the transition between soluble and aggregated states of the protein. The data illustrates the overlap between the native and amyloid propensities in the protein sequences, contributing to explain why proteins cannot avoid the presence of aggregation-prone regions.

INTRODUCTION

The function of a large majority of polypeptides depends on the attainment of a globular, compact and specific three-dimensional structure after their synthesis at the ribosomes [1]. Only properly folded globular proteins can interact specifically with their molecular targets [2]. The protein quality machinery works to minimize the accumulation of misfolded species, not only because they are not functional, but also because these conformers often display an intrinsic propensity to self-assemble into toxic aggregates, provoking the impairment of essential cellular processes. Accordingly, protein misfolding and aggregation lie behind an increasing number of human diseases that include highly debilitating disorders like Alzheimer's or Parkinson's disease [3]. Despite the polypeptides causing these pathologies are not related in terms of sequence or structure, in many cases their aggregation leads to the formation of amyloid fibrils, all sharing a common cross-β motif [4]. However, the adoption of amyloid-like conformations is not restricted to disease-linked proteins and might constitute a generic property of polypeptide chains [5,6,7], likely because the non-covalent contacts that stabilize native structures resemble those leading to the formation of amyloids [8].

It was initially thought that, for globular proteins not linked to disease, amyloid fibril formation involved the docking of monomeric partially folded states, which display pre-existent β-sheet structure. Nevertheless, it was early shown that all-α proteins can also be induced to form amyloids under strongly destabilizing conditions. In particular, Dobson and co-workers demonstrated that in the case of apomyoglobin, amyloid fibril formation correlates with environments in which the protein backbone is unfolded, rather than with conditions that may allow population of partially structured states enriched in β-sheet conformations [9,10]. Destabilization of apomyoglobin by mutation of two highly conserved Trp residues to Phe also results in the formation of amyloid fibrils, under conditions close to physiological [11]. For this double mutant, solution conditions that promote the population of the native α-helical secondary structure abolish the polymerization of the protein [11], illustrating a competition between folding and aggregation. In the present work we use the FF domain to provide further insights into the mechanism of amyloid fibril formation by α-helical proteins.

FF domains are small protein-protein interaction modules consisting of ~50-70 residues often organized in tandem arrays and characterized by the presence of two conserved Phe residues at the N- and C-termini [12]. The three-dimensional structures

of several FF domains have been solved, showing that this fold consists of three α-helices arranged as an orthogonal bundle with a 3_{10} helix in the loop connecting the second and the third helix [13,14,15]. They are involved in RNA splicing, signal transduction and transcription processes [16,17]. These domains are present in a variety of eukaryotic nuclear transcription and splicing factors as well as in p190RhoGTPase-related proteins and their sequences are well conserved from yeast to humans [12]. However, the sequences of different FF domains are highly divergent. The loops connecting the different α-helical regions display the highest sequential variability, both in length and amino acidic composition. The main structural difference between divergent FF domains is the orientation, and sequence, of the second helix, which has been proposed as the structural element responsible for ligand specificity.

The folding process of the FF domain from human HYPA/FBP11 (HYPA/FBP11-FF) has being characterized in detail [18,19,20,21]. This domain is receiving considerable attention, especially because it forms early in the folding reaction an on-pathway, short-lived and low populated intermediate state whose structure has been solved at atomic resolution combining NMR relaxation dispersion methods and computational techniques, providing thus a molecular description of a transient folding intermediate with unprecedented detail [22,23,24,25].

The aggregation properties of FF domains have not been characterized yet. Here, we address this issue using the yeast URN1 FF domain (URN1-FF) as a model system. This is the only FF domain of the yeast URN1 protein consisting of 59 residues and adopts a canonical α1-α2-3_{10}-α3 FF fold (Fig 1). Relative to HYPA/FBP11-FF, the three α-helices are shorter in the yeast domain. Also, helices 1 and 3 are closer and more orthogonal in URN1-FF. We show here that URN1-FF forms amyloid fibrils at low pH. Using a battery of biophysical techniques to study the conformational, thermodynamic and kinetic properties of soluble URN1-FF, the morphological, structural and tinctorial properties of its aggregates as well as dissecting this domain into its individual secondary structure elements, we demonstrate that helix 1 at the N-terminus plays a key role in controlling the conformation and solubility of this small all-α protein.

RESULTS

pH dependence of URN1-FF conformational properties

In contrast to HYPA/FBP11-FF, which displays a basic pI of 9.9, the pI of URN1-FF is 4.4. We studied the conformational properties of URN1-FF over the pH range 2.0–5.7 and 298 K by far-UV circular dichroism (CD), intrinsic fluorescence and 1-anilinonaphtalene-8-sulfonate (ANS) binding at 20 μM protein concentration (Fig 2).

The far-UV CD of URN1-FF at pH 5.7 is similar to that reported for HYPA/FBP11-FF in the same conditions and, in agreement with its solution NMR structure, it corresponds to that of a canonical α-helical protein, displaying the characteristic minima at 210 and 222 nm (Fig 2A). Analysis of the CD spectra with the K2D3 algorithm predicts a secondary structure content consisting of 95% of α-helix and 5% or random coil without any significant β-sheet component. The far-UV CD spectra of URN1-FF solutions at different acidic pHs (2.0, 2.5 and 3.0) all display an α-helical spectra similar to that under native conditions at pH 5.7, despite exhibiting reduced ellipticity at pH 2.0 (Fig 2A), suggesting that this domain retains a significant amount of their native secondary structure at low pH. K2D3 predicted an α-helix content of 95%, 95% and 93% at pH 3.0, 2.5 and 2.0, respectively, without any significant contributions of β-sheet signal. In contrast to the other conditions, the protein solution becomes cloudy at pH 4.0. This was probably because this pH is close to the URN1-FF pI, resulting in isoelectric protein precipitation. Therefore this condition was not further analyzed.

URN1-FF contains two buried Trp residues at positions 27 and 56. We monitored the pH-induced changes in the tertiary structure of this domain following the variation in Trp emission. A progressive increase in fluorescence intensity and a red shift of the maximum emission wavelength was observed as pH decreases, indicating an increase in the solvent exposure of these aromatic residues (Fig 2B). Importantly, at the lowest pH value tested here the maximum emission was ~335-340 nm, indicating that the Trp residues are not fully solvent-exposed. This suggested opening of the native conformation with the concomitant partial exposure of hydrophobic clusters at low pH. To test this possibility further, we analyzed the binding of URN1-FF to ANS at the different pHs. At all pH values the protein was found to increase the fluorescence of ANS and cause a blue-shift of its emission maximum. However, a dramatic increase in

ANS fluorescence was observed at pH 2.0 and 2.5, with intermediate and low ANS fluorescence increases at pH 3.0 and 5.7, respectively (Fig 2C).

The structural features of the URN1-FF at pH 2.5 and 5.7 were also evaluated by NMR spectroscopy. The one-dimensional NMR (^{1}H-NMR) spectrum of the protein at pH 5.7 displayed a wide signal dispersion of resonances at both low (amide and aromatic region) and high (methyl region) fields, with good peak sharpness, characteristic of folded molecules (Fig 2D). At pH 2.5, certain peak collapse was observed. However, the signal dispersion is indicative of the retention of a significant number of URN1-FF intramolecular contacts at this pH.

Overall, the structural features of URN1-FF at acidic pH are compatible with the population of a molten globule-like structure in these conditions. To assess the stability of the detected URN1-FF species, we monitored the evolution of their conformational properties in a wide range of pH, from 1.5 to 6.5 along time. In the 2.0-6.5 pH range no changes in tertiary structure, as monitored by Trp fluorescence, and secondary structure, as monitored by CD, were detectable upon 24 h incubation. Accordingly, no aggregation signs were visible by light scattering, in this time period and pH range (Fig S1) indicating that the conformational ensembles are at least metastable.

pH dependence of URN1-FF thermal and chemical stability

The thermal stability of URN1-FF under native conditions (pH 5.7) was analyzed following the changes in the far-UV CD spectra at 222 nm and in intrinsic fluorescence at 350 nm (Fig 3A and 3B, respectively). The two probes reported essentially identical thermal transitions, with a melting temperature T_m of 340.9 ± 0.3and 341.4 ± 0.1 K, by far-UV CD and intrinsic fluorescence, respectively. The data were therefore analyzed according to a two-state model for unfolding (Table 1). We addressed then the dependence of the thermal stability of the domain on the pH. At all assayed pHs, cooperative transitions that could be analyzed according to a two-state thermal unfolding model where observed. Nevertheless, the thermal stability of URN1-FF is drastically affected by the pH. The T_m decreases with decreasing pH and is reduced by ~30 K from pH 5.7 to 2.0 (Table 1). A good agreement between the thermodynamic values calculated from CD and fluorescence data is observed at all pHs.

To analyze differences in the stability of the different URN1-FF conformations, we studied the resistance of the protein at different pH values against chemical denaturation with urea at 310 K by monitoring the changes in molar ellipticity at 222

nm and in Trp fluorescence intensity at 350 nm (Fig 3C and 3D, respectively). We selected 310 K because at this temperature the percentage of folded URN1-FF strongly depends on the pH. As it will be seen in the next sections, this will allow us correlate the degrees of native structure and aggregation. At pH 5.7 the FF domain unfolded in a cooperative, two-state process. Accordingly, the thermodynamic values obtained from fluorescence and CD measurements were similar (Table 2). ΔG^{H2O} of 5.33 ± 0.25 kcal/mol and 5.94 ± 0.18 kcal/mol were calculated by far-UV CD and intrinsic fluorescence, respectively. A $[urea]_{50\%}$ of 6.75 M was obtained for both probes. This domain is significantly more stable than the structurally homologous human HYPA/FBP11 FF domain, which has a ΔG^{H2O} of 3.6 kcal/mol and a $[urea]_{50\%}$ of 3.1 M at the same pH.

A dramatic decrease of URN1-FF thermodynamic stability was observed at lower pHs. At pH 3.0 and 2.5 the protein exhibited a complete cooperative transition. CD and fluorescence analysis rendered similar thermodynamic parameters showing that the protein is destabilized by ~ 3.6 and ~ 4.4 kcal/mol at pH 3.0 and 2.5, respectively (Table 2). As expected, at pH 2.0 the domain is partially unfolded and displays marginal stability since 310 K is close to the T_m at this pH.

Folding and unfolding kinetics of the URN1-FF domain

The kinetics of folding and unfolding of the URN1-FF domain at pH 5.7 were determined by stopped-flow and fluorescence detection under a wide range of denaturant conditions at 298 and 310 K. In both cases, the folding and unfolding traces by fluorescence fit well into single exponentials. Moreover, the chevron plots appear to be linear in the complete range of urea concentration studied (Fig 4A), indicating the lack of detectable intermediates, according with a two-state model. The rate constants for folding (k_f) and unfolding (k_u) and their dependence on denaturant concentration (m_f and m_u) are shown in Table 3. The thermodynamic data at 310 K obtained from kinetic measurements are in good agreement with the equilibrium data at this temperature with chemical denaturation. Increasing the temperature from 298 to 310 K has a moderate effect on the thermodynamic stability of URN1-FF (~ 0.3 kcal/mol), but it has an important impact on the folding and unfolding rates, which increase by 2.6 and 5.8 fold, respectively. The folding and unfolding kinetics of the HYPA/FBP1 domain have been characterized at 283 K. Unfortunately, URN1-FF could not be studied at this temperature since at the high concentrations required for unfolding reactions urea

crystallizes, precluding direct comparison of kinetic traces. However, it is worth mentioning that even at 310 K the unfolding rate of URN1-FF is two fold slower than that of HYPA/FBP1-FF at 283 K, suggesting that the unfolding rate is a main contributor to the different thermodynamic stabilities exhibited by these two FF domains.

The pH dependence of the unfolding rate at 8 M urea was measured at 310 K over the pH range 3.5-6.5 (Fig 4B). Below pH 3.5 the unfolding reaction was too fast to be monitored with our equipment. As it happens with HYPA/FBP1-FF, lowering the pH results in an increase in the k_u value for URN1-FF. The plot of ln (k_u) versus pH allows approximating the unfolding rate constants at pH 2.5 and pH 2.0, 8M urea and 310 K by extrapolation. Then, assuming that m_u is independent of pH, $k_u[H_2O]$ values of ~ 150 and ~ 360 s^{-1} are calculated at pH 2.5 and 2.0, respectively. This indicates that the domain unfolding is very fast under these conditions.

Effect of pH on the aggregation of the URN1-FF domain

The population of molten-globule like states has been correlated with the formation of amyloid aggregates [26]. Despite URN1-FF remains soluble at 20 μM at all pHs, with the exception of pH 4, it might aggregate at higher concentrations. To assess if the URN1-FF domain self-assembles into macromolecular structures in a pH dependent manner, the protein was incubated at 140 μM and 310 K for 7 days over the pH range 2.0–5.7. The presence and morphology of protein aggregates was analyzed using Transmission Electron Microscopy (TEM) (Fig 5A). Fibrillar species were observed at both pH 2.0 and pH 2.5. The fibrils in both solutions were long and unbranched and consist of linear fibrils with a diameter of ~7 nm or twisted fibrils with a diameter of ~14 nm. At pH 3.0 only small protofibrillar-like aggregates were observable and at pH 5.7 the solution was essentially devoid of aggregates. Quantification of the amount of aggregated protein by sample fractionation using sedimentation at 100,000*g* for 1 hour is consistent with the TEM analysis, showing that most of URN1-FF is aggregated at pH 2.0, 2.5 and 3.0, with 97%, 99% and 100% of the total protein being located in the insoluble fraction, respectively. On the contrary, 93% of the protein remained in the soluble fraction at pH 5.7. As expected, the protein solution became immediately cloudy at pH 4.0. Accordingly, large and amorphous aggregates were observed by TEM (Fig S2).

We used the amyloid-specific dyes thioflavin T (ThT) and Congo red (CR) to analyze if the aggregates detected in the different protein solutions exhibit amyloid-like features (Fig 5B and 5C). In agreement with their fibrillar appearance, the aggregates formed at pH 2.0 and 2.5 promoted the highest increase in ThT fluorescence emission, followed by the pH 3.0 protofibrillar assemblies; little ThT binding was observed at pH 5.7. With the exception of pH 5.7, incubation of URN1-FF at all other pHs promoted binding of the protein to CR resulting in a red-shift and an increase in the absorbance maximum of the dye, thus confirming the presence of amyloid-like aggregates formed in these conditions.

We then monitored the conformational properties of the URN1-FF aggregates incubated in the different acidic conditions using far-UV CD (Fig 5D). The protein incubated at 140 μM, pH 5.7, 310 K for 7 days exhibited essentially the same native spectrum as the freshly dissolved protein at 20 μM, displaying the characteristic minima at 210 and 222 nm (Fig 5D). In contrast, in FF domains incubated at pH 2.0, 2.5 and 3.0 the transition towards a β-sheet enriched conformation was evident.

We also monitored the exposure of hydrophobic clusters to the solvent in the aggregates attained at the different pHs with ANS (Fig 5E). The fibrils formed at pH 2.0 exhibited the highest ANS binding, suggesting conformational differences between these assemblies and the fibrils formed at pH 2.5, in line with their different binding to CR and β-sheet signal intensity in the CD spectra.

Overall, the data converge to indicate that the URN1-FF domain forms aggregates displaying different morphological and conformational properties depending on the pH. In particular, the aggregates at pH 2.0 and 2.5 have the morphology, structural and tinctorial characteristics typical of amyloid fibrils.

The aggregation kinetics of the URN1-FF domain at low pH

The time course of URN1-FF amyloid fibril formation at 140 μM, pH 2.5 and 310 K was monitored by light scattering and ThT fluorescence. The kinetics of amyloid fibril formation usually follows a sigmoidal curve that reflects a nucleation-dependent growth mechanism. The assembly of URN1-FF follows this kinetic scheme but exhibits a short lag phase (Fig 6 and Table S1). The kinetic traces, as well as the lag time and the elongation rate constant for the aggregation reaction, followed by ThT binding and light

scattering were essentially identical, indicating that at pH 2.5, the formation of amyloid-like intermolecular interactions occurs rapidly in the aggregation process.

Prediction of URN1-FF sequence segments with α-helical and aggregation propensities

It is now accepted that specific and continuous sequence segments of a protein promote amyloid-like reactions and participate to the formation of the β-core of the mature fibrils [27,28]. Different computational methods have been developed to predict those sequential stretches [29,30]. The evidence that at low pH URN1-FF forms amyloid fibrils indicates that this domain should posses at least one amyloidogenic region. To this purpose we have used a number of the existing algorithms to identify the region of the URN1-FF sequence that most likely promote amyloid fibril formation of the protein. WALTZ detects the stretch spanning residues 12-17 in helix 1 at pH 2.5. Consistently, PASTA detects a single amyloidogenic region comprising residues 10-16. In addition to helix 1, BETASCAN and FOLDAMYLOID algorithms predict a second sequence comprising residues 25-20 and 25-29, respectively, at the end of loop 1 and beginning of helix 2. Finally, in addition two these two stretches, AGGRESCAN suggests that residues 50-55 in helix 3 may also display a moderate aggregation propensity (Fig 7A).

The predicted α-helical propensity of the URN1-FF sequence was analyzed using AGADIR in the 2.4-5.8 pH range at 310 K. The global α-helical propensity of the protein is predicted to decrease with decreasing pH. At pH 5.7 the region of highest α-helical propensity corresponds to residues 6-21, including the complete helix 1 and the initial part of loop 1, the rest of residues display low intrinsic α-helical propensity. Helix 1 is also the region with the highest α-helical propensity at pH 2.5 but its predicted α-helical propensity is two fold lower.

To confirm the above conformational and aggregation predictions we designed five peptides encompassing the entire URN1-FF amino acid sequence (Fig 7B). The N_t peptide corresponds to residues 1-7 at the N-terminus, which are devoid of any regular secondary structure in the native protein. The H1 peptide corresponds to the complete helix 1 (residues 8-20). The H2 peptide comprises the stretch 21-35 including loop 1 and helix 2. The 3_{10} peptide includes loop 2 and the 3_{10} helix. These residues (36-45) connect the helices 2 and 3 in the native structure. Finally, the H3 peptide includes

residues 46-59 at the C-terminal helix 3. All peptides were prepared with both unprotected termini.

Conformational properties of synthetic URN1-FF peptides

We first assessed the conformational properties of the URN1-FF peptides using far-UV CD at 100 μM, in 100 mM glycine buffer at pH 2.5 and 298 K (Fig 8A-E). In this condition all the peptides exhibited an essentially unstructured conformation with minima below 200 nm, in agreement with the general observation that α-helices are only slightly stable in most short peptides in aqueous solution [31,32]. Trifluoroethanol (TFE) has been recurrently used to stabilize the α-helical structure in short peptides [33,34,35]. Several proteins have been split into peptide fragments which showed a tendency to form α-helices in TFE, even if they were unstructured in water. This propensity is particularly strong for peptides corresponding to α-helical regions in the intact protein [36]. Therefore, we tested the effect of TFE in the 5-25% (v/v) concentration range on the CD spectra of URN1-FF peptides at pH 2.5 (Fig 8A-E). In agreement with the poor secondary structure content of the correspondent protein regions in the native state, the N_t and 3_{10} peptides did not show evidence of secondary structure formation in any solvent condition. The H1 and H3 peptides exhibited a shift of the molar ellipticity minima toward 210 at 222 nm at increasing TFE concentrations indicating the adoption of an α-helical conformation, consistent with these two segments being the largest α-helical regions in the native URN1-FF structure. The molar ellipticity at 222 nm also increases with TFE concentration in the H2 peptide, despite the signal at 210 nm is not evident, indicating a certain propensity to populate the α-helical state but clearly lower than in the case of the H1 and H3 peptides. This is consistent with a smaller length of helix 2 in the native state. The plot of molar ellipticity at 222 nm versus TFE concentration allows comparing the effect of the solvent on the gain of α-helical conformation in the different peptides (Fig 8F). The TFE effect is higher for H1, followed by H3 and then H2, with no apparent structural induction in the N_t and 3_{10} peptides. These results are in agreement with the highest α-helical propensity predicted by AGADIR for helix 1 at pH 2.5.

Aggregation properties of synthetic URN1-FF peptides

As a next step, we analyzed the properties of the URN1-FF peptides upon incubation under aggregation-promoting conditions (1500 μM for the N_t peptide and 600 μM for the rest of the peptides, in 100 mM glycine buffer at pH 2.5 and 310 K) in the absence and in the presence of 15% and 25% (v/v) of TFE using far-UV CD and compared them with the ones exhibited by the peptides under non-aggregating conditions (100 μM, 100 mM glycine buffer, pH 2.5 and 298 K) (Fig 9A-E). A negative peak at ~217 nm, indicative of β-sheet structure, was observed in the CD spectrum of the H1 peptide in the absence of TFE. The β-sheet signature is still evident in the presence of 15% (v/v) of TFE. In contrast, when the H1 peptide was incubated in the presence of 25% (v/v) TFE it adopted a random coil conformation and the spectrum overlapped with the one of the peptide under non-aggregating conditions in the absence of the alcohol. We monitored the morphology and amyloid-like features of the peptide aggregates by ThT fluorescence (Fig 9F) and TEM binding (Fig 9G-K) in the absence and presence of 25% (v/v) TFE. In the absence of TFE, fibrillar structures displaying a strong binding to ThT were observed, compatible with an amyloid-like structure. The presence of TFE reduced drastically the formation of aggregates and abolished ThT binding.

The H3 peptide formed amorphous aggregates under aggregating conditions in the absence of TFE, as imaged by TEM, causing a significant loss of the far-UV CD signal, relative to that observed under non-aggregating conditions. Accordingly, these aggregates exhibited low binding to ThT. In contrast to H1, the presence of 25% (v/v) TFE promoted the appearance of a β-sheet signal in the CD spectrum of H3, an increase in ThT fluorescence, and the formation of smaller peptide aggregates, as observed by TEM.

The N_t, H2 and 3_{10} peptides exhibited predominantly random coil structure under aggregating conditions in the absence of TFE. These three peptides were insensitive to the presence of TFE: small changes were observed in their far-UV CD spectra, little aggregation was visualized by TEM and no significant binding to ThT was observed, indicating that they remain essentially soluble even at high concentration and low pH.

Accessibility of α-helices in the URN1-FF domain at low pH

The peptide analysis suggests that helix 1 displays the highest α-helical and aggregation propensities at low pH and that marked propensities are also displayed by the H3 peptide, albeit to a lower extent. We used limited proteolysis to test if these two α-helices display detectable flexibility at 140 μM, 100 mM glycine buffer, pH 2.5 and 310 K in the complete URN1-FF domain and can therefore be cleaved by proteases. The FF domain was incubated with pepsin and the progress of the digestion reaction was monitored by SDS-PAGE on Tricine gels and by MALDI-TOF mass spectrometry (MS). A major band appeared after 30 seconds of digestion (Fig 10A) with a molecular weight of 4587 Da (Fig 10B), which could be assigned to residues 16-52. After 1 minute, a main fragment of 3108 Da was detected. This fragment corresponds to residues 16-40 and remains resistant to proteolytic attack even after prolonged incubation. It results from an internal cleavage of the original 16-52 segment. Accordingly, the appearance of the 3108 Da peak in the MS spectrum is always associated with the presence of a 1495 Da peak that matches with the predicted mass of the 41-52 fragment. Therefore, the data are consistent with the first rapid cleavages taking place inside helices 1 and 3 (Fig 10C), suggesting that they display a high conformational flexibility and low protection in the monomeric molten-globule state of the protein at low pH.

DISCUSSION

It is now accepted that amyloid fibrillation constitutes a generic property of polypeptide chains, including globular proteins [1]. Regardless of the amino acid sequence, amyloid fibrils comprise a common cross-β structure [4]. The corollary is that the competition between native and cross-β conformations is inherent to most globular proteins. The native structure has evolved under natural selection and is sustained by specific interactions between side-chains, which determine the backbone conformation. By contrast, repetitive backbone-backbone interactions dominate amyloid fibrils, with side chains being accommodated in the most favourable disposition compatible with the cross-β structure. Nevertheless, side chain contacts are crucial determinants of the initial transition of soluble proteins towards amyloid states, modulating thus amyloid propensity [27]. Similarly, the propensity of a polypeptide to form specific secondary structures is assumed to depend largely on its primary sequence. The transition of native α-helices to amyloid β-sheets illustrates how these structural propensities might overlap and are modulated by the structural context. α-helices and β-sheets represent alternative ways of saturating all the hydrogen bond donors and acceptors in a polypeptide backbone. Therefore, fluctuations between these two types of secondary structures involve the disruption and establishment of a significant number of non-covalent interactions. Understanding the transition between these two conformations is of interest since it has been suggested to underlie aggregation of the Aβ peptide and α-synuclein in Alzheimer's and Parkinson's diseases, respectively [37,38].

Destabilization of the native state, either by solution conditions or by mutations, is a usual requirement for amyloid fibril formation in globular proteins. It usually promotes the population of partially unfolded conformers that otherwise are inaccessible or in fast equilibrium with the native state. Mild denaturation at low pH has been used in different protein models to induce amyloid fibril formation [39,40]. Here, we have studied the aggregation propensity of the URN1-FF domain as a function of the pH. We observed that for this all-α protein, the α-helical isomers populated below pH 3.0 display a high propensity to experiment a conformational transition leading to the formation of β-sheet enriched amyloid fibrils at physiological temperature. A variety of methodologies probing the conformation of soluble and aggregated states of URN1-FF and its secondary structure elements have been employed to gain insights into the mechanistic aspects of the fibrillation reaction.

URN1-FF remained soluble at 20 μM in the pH range 2.0-6.5 for 24h. The different isomers formed in these conditions maintained their conformational properties in this time window indicating that they are at least metastable. Importantly, no significant population of β-sheet structures was detected at this protein concentration in the complete pH range. Thermal denaturation at pH 2.0-3.0 render cooperative transitions with similar T_m and traces for CD and fluorescence probes indicating that despite the different secondary and tertiary structure content in the conformers populated at the different acidic pHs all them appear to unfold following a two-state model, without the apparent population of intermediates. At pH 3.0 and 2.5 the protein keeps essentially the same α-helical content than the native protein, however thermal denaturation indicates that the protein is destabilized in these conditions. At pH 2.0, the α-helical content decreases slightly and the protein stability is significantly reduced. Overall, the species populated at low pH have characteristics compatible with molten globule-like conformations.

As in the case of other model globular proteins, solution conditions promoting the aggregation of this α-helical domain coincide with those allowing the presence of a significant amount of native secondary structure since the types of interactions sustaining both types of structures are essentially identical. As expected, the high stability of the URN1-FF native structure at pH 5.7 prevents the protein from aggregation in this condition. Fitting of the chemical denaturation data to a two-state folding model indicates that, at 310 K, ~50% of the URN1-FF molecules are unfolded at pH 2.0 and equilibrium. At pH 2.5 the unfolded population decreases to 10%, whereas at pH 3.0 the protein is essentially in a molten-globule state with native-like secondary structure. Therefore, the population of a large amount of unfolded species at equilibrium is not a requirement for URN1-FF aggregation into β-sheet enriched aggregates. CD and fluorescence probes indicate that lowering the pH promotes weakening of the hydrogen bonds sustaining the URN1-FF α-helical structure with a concomitant loss of tertiary contacts. At low concentration, the main conformational difference between pH 3.0 and 5.7 protein solutions is the exposition of hydrophobic residues to solvent at pH 3.0. This effect is even more dramatic at lower pHs. It is likely that URN1-FF aggregation will be initiated by the establishment of intermolecular contacts between hydrophobic residues in these partially folded conformers that were previously hidden in the native structure.

At pH 2.5, the aggregation rate is several orders of magnitude slower than the rate for conformational unfolding. This suggests that whereas exposure of hydrophobic residues to solvent suffices to induce intermolecular self-association at elevated concentration, the conformational conversion from α-helix to amyloid structures requires a significant destabilization of the protein and at least partial unfolding of the native α-helical structure; likely because the formation of ordered intermolecular hydrogen bonds between β-sheets is effectively competed by URN1-FF native helices. This will explain why ordered amyloid fibrils are observed at pH 2.5 and not at pH 3.0, where the protein is more stable, reducing the probability of forming extended conformations that can be aligned into an array of β-sheets. Accordingly, proteolysis experiments suggest that at pH 2.5 the α-helices display strong conformational flexibility at this pH. The coincident kinetics for ThT and light scattering at pH 2.5, together with the absence of amorphous aggregates in EM images, indicate that fibrils are preferentially formed at this pH and that if protofibrillar assemblies are populated in this condition, they rapidly convert to fibrillar species, a reaction that does not occur or occurs very slowly at pH 3.0.

Computational predictions pointed out the helix 1 as the URN1-FF sequence stretch with both the highest α-helical and amyloid propensities. Experimental analysis of peptides corresponding to the individual secondary structure elements of URN1-FF confirmed this extent. Moreover, characterization of the transition state for folding of the structurally homologous HYPA/FBP11-FF domain indicates that the helix 1 is part of the folding nucleus [41] and this α-helix is already formed in an early intermediate in the folding pathway of this domain [42]. Kinetic partitioning of protein folding and aggregation reactions has been described for some proteins, in such a way that regions of the sequence responsible for initiating the process of aggregation do not participate in the establishment of the folding nucleus [43]. This is likely not the case for URN1-FF. This domain illustrates how nature finds difficult to avoid the presence of aggregation-prone regions in proteins because, at least in certain cases, residues leading to the formation of native structures are also able to trigger self-assembly into toxic conformers since both processes involve the development of similar inter-residues interactions between transiently unfolded protein regions. In these cases, it is the stability of the native conformation, and more specifically the protection of the sequence stretches with the highest aggregation propensities, that prevents the transition

towards amyloidogenic conformations. Accordingly, natural mutations associated with familial forms of amyloidosis have been shown to reduce the stability of the folded state [44,45]. As illustrated here for the helix 1, stabilization of hydrogen bonds in α-helical structures, in this case by adding TFE, reduces their ability to self-assemble into β-sheets. Therefore, compounds that target α-helical structure in disease-related proteins and reduce their conformational fluctuation might become specific aggregation inhibitors. In fact, this approach has already proven successful to reduce Aβ peptide induced *in vivo* neurodegeneration [37] and has been suggested as a promising strategy to tackle α-synuclein aggregation in Parkinson's disease, dementia with Lewy bodies and other synucleinopathies [38]. Overall, the FF domain emerges as a new and useful protein model to dissect the molecular determinants accounting for the transition between soluble and aggregated protein states.

MATERIALS AND METHODS

Protein expression and purification

The URN1-FF domain corresponds to residues 212-266 of yeast URN1 and was cloned into a pETM-30 vector as an N-terminal fusion protein with a His tag followed by GST and a TEV protease cleavage site [15]. For protein production, the plasmid was transformed in BL21 (DE3) cells and after growing to 0.6 optical density they were induced with 1 mM IPTG at 298 K overnight. As a first step of purification, a His-tag column was used to isolate the GST-fused protein. Subsequently, a TEV cleavage was performed and a final gel filtration HiLoad™ Superdex™ 75 prepgrade (GE healthcare Life Sciences) was used to remove the GST protein. The sample was dialyzed against water and lyophilized. The purity of the samples was checked by SDS-PAGE and MALDI-TOFF mass spectroscopy. Protein concentration was determined by UV absorption using a ε value of 1.948 mg^{-1} ml cm^{-1}.

Sample preparation for URN1-FF and synthetic peptides conformational assays

Lyophilized URN1-FF protein was dissolved at 20 μM using a wide range of pH solutions: 50 mM potassium chloride at pH 1.5 and 1.75; 50 mM glycine at pH 2.0, 2.25, 2.5, 2.75 and 3.0; 50 mM sodium acetate at pH 3.5, 4.0, 5.0, 5.5 and 5.7; and 50 mM MES at pH 6.0 and 6.5. Protein solutions were filtered through a 0.22 μm filter and immediately analyzed at 298 K by Tryptophan intrinsic fluorescence, far-UV CD, static light scattering and ANS binding. Synthetic peptides were dissolved in 100 mM glycine at pH 2.5 and sonicated during 10 minutes. In all the cases the final concentration was 100 μM and different amounts of TFE were added between 0 and 25% (v/v).

Sample preparation for URN1-FF and synthetic peptides aggregation assays

Lyophilized URN1-FF protein was dissolved at 140 μM in different pH buffers and filtered through a 0.22 μm filter. Five different pH conditions were chosen to check aggregation: 100 mM glycine at pH 2.0, 2.5 and 3.0; and 100 mM sodium acetate at pH 4.0 and 5.7. Lyophilized synthetic peptides were prepared at 1500 μM for the N_t peptide and 600 μM for the rest of the peptides in 100 mM glycine at pH 2.5 in the presence of 0, 15 and 25% (v/v) of TFE and sonicated during 10 minutes. In all the cases, the samples were incubated under agitation at 400 rpm and 310 K during 7 days.

Aggregation kinetics

URN1-FF protein was prepared at 140 μM in 100 mM glycine at pH 2.5 in the presence of 25 μM of ThT. Immediately after equilibrating the sample at 310 K during 5 minutes, ThT intrinsic fluorescence and light scattering intensity were measured every 5 minutes during 2000 minutes. The sample was excited at 440 nm and emission measurements were recorded at 475 nm for ThT intensity. For light scattering intensity, the excitation and emission wavelengths were at 340 and 350 nm, respectively. Slit widths of 5 nm were used for both excitation and emission in a CARY-100 Varian spectrophotometer.

Electron Microscopy

Samples were diluted tenfold with water and 10 μl were placed on carbon-coated copper grids and left for 5 min. The grids were then washed with distillated water and stained with 2% (w/v) uranyl acetate for 1 min. The analysis was done using a HITACHI H-7000 transmission electron microscope operating at an accelerating voltage of 75 kV.

Binding to amyloid dyes

URN1-FF aggregates were diluted at 10 μM in phosphate buffer pH 7.5 containing 25 μM of ThT. Synthetic peptides were studied at 40 μM with the same amount of ThT. ThT was excited at 440 nm and fluorescence emission was recorded between 460 and 600 nm, using excitation and emission slit widths of 10 nm. Each trace was the average of 3 accumulated spectra at 298 K in a CARY-100 Varian spectrophotometer.

To study the binding to Congo red, 30 μl of aggregated URN1- were mixed with 220 μl of CR (20 μM) in 5 mM phosphate, 150 mM NaCl pH 7.4 buffer at 298 K. After 5 minutes of equilibration, optical absorption spectra were acquired from 400 to 700 nm and accumulated for 3 times with a Jasco V-630 spectrophotometer (Tokyo, Japan). Solutions containing only protein and only CR were analyzed to eliminate the protein scattering and dye contribution.

ANS binding assay

Aggregated samples were diluted at 10 μM in phosphate buffer pH 7.5 containing 25 μM of ANS. To study soluble URN1-FF species, samples were prepared

at 20 μM containing 25 μM of ANS and analyzed immediately. The excitation wavelength was 370 nm and the emission spectra was recorded between 400 and 600 nm, using excitation and emission slit widths of 5 and 10 nm, respectively. Three spectra were accumulated after 5 minutes of equilibration at 298 K in a CARY-100 Varian spectrophotometer.

NMR spectroscopy

Lyophilized protein was dissolved at 35 μM in water using a 9:1 H_2O/D_2O ratio, and adjusted at pH 2.5 and pH 5.7. One-dimensional NMR spectra were acquired at 298 K on a Bruker AVANCE 600-MHz spectrometer using solvent suppression WATERGATE techniques.

Circular dichroism, tryptophan intrinsic fluorescence and static light scattering

Monomeric and aggregated URN1-FF species were prepared at 20 μM and measured immediately. Far-UV CD spectra were measured in a Jasco-710 spectropolarimeter thermostated at 298 K. Spectra were recorded from 260 to 200 nm, at 0.2 nm intervals, 1 nm bandwidth, and a scan speed of 50 nm/min. Twenty accumulations were averaged for each spectrum. Both monomeric and aggregated synthetic peptides were diluted to a final concentration of 100 μM. Far-UV CD spectra were recorded in a Jasco J-810 spectropolarimeter (Tokyo, Japan) thermostated at 298 K, using the same conditions described above.

Tryptophan intrinsic fluorescence was measured at 298 K in a Cary-100 Varian spectrofluorometer using an excitation wavelength of 280 nm and recording the emission from 300 to 400 nm. Three averaged spectra were acquired and slit widths were typically 5 nm for excitation and emission. Soluble URN1-FF samples were measured in a PerkinElmer Life Sciences LS-55 fluorimeter (Wellesley, Massachusetts) equipped with a thermostated cell compartment using the same parameters described above. The represented values are the integral of the fluorescence between 300 and 400 nm.

Static light scattering was recorded in a Jasco V-630 spectrophotometer (Tokyo, Japan) at 298 K. Two accumulative spectra were registered between 360 and 240 nm.

Thermal denaturation

URN1-FF samples were dissolved at 20 μM and four different pH conditions (2.0, 2.5, 3.0 and 5.7). Thermal stabilities were studied by intrinsic fluorescence intensity and far-UV CD. The samples were excited at 280 nm and the emission was recorded at 360 nm, using slit widths of 5 nm for excitation and emission in a CARY-100 Varian spectrophotometer. The measurements were registered each 0.25 K with a rate of 0.5 K/min. The molar ellipticity at 222 nm was recorded each 0.2 K with a rate of 0.5 K/min in a Jasco-710 spectropolarimeter. Experimental data were fitted to a two-state transition curve using the Kaleidagraph version 4.0 (Synergy Software).

Equilibrium stability measurements

Lyophilized protein was dissolved at 20 μM in different pH buffers and molarities of urea, from 0 to 9 M and left to equilibrate for at least 3h. The samples were equilibrated during 10 minutes at 310 K and analyzed by intrinsic fluorescence intensity and far-UV CD. The excitation and emission wavelengths were 280 and 360 nm, respectively, using a CARY-100 Varian spectrophotometer. Chemical unfolding was followed by far-UV CD at 222 nm in a Jasco-710 spectropolarimeter. Experimental data were fitted to equation assuming two-state unfolding using the Kaleidagraph version 4.0 (Synergy Software).

Folding kinetics

Kinetics of unfolding and refolding reactions at 298 K and 310 K were followed in a Bio-Logic SFM-3 stopped-flow instrument using excitation at 280 nm and a 320 nm fluorescence cut-off filter. Unfolding reaction was promoted by dilution of the protein in buffer with appropriate volumes of the same buffer containing 9.5 M urea. For folding reactions, appropriate volumes of buffer were added to an initial protein solution containing 9.5 M urea. To determine the unfolding rates at 8 M of urea, protein and urea solutions were prepared at different pH and unfolding kinetics were recorded 310 K in a Bio-Logic SFM-3 device.

Observed rate constants were fitted to the equation describing folding of a two-state protein, using the Kaleidagraph version 4.0 (Synergy Software). To determine the free energy, m_{eq} and C_m values we used the following equations:

$$\Delta G_{F-U} = RT \ln(k_F / k_U)$$

$$m_{eq} = RT(m_f + m_u)$$

$$C_m = \Delta G_{F-U} / m_{eq}$$

where k_f and k_u are the rate constants for folding and unfolding, respectively, and the m_f and m_u values correspond to the slopes of the respective folding and unfolding regions.

Prediction of the aggregation-prone regions and α-helical propensity of URN1-FF domain

The primary sequence of URN1-FF was used as input to predict the regions prone to aggregation. We used five different algorithms: WALTZ [46], AGGRESCAN [47], BETASCAN [48], FOLDAMYLOID [49] and PASTA [50]. The tendency of URN1-FF to adopt α-helical structure was predicted using AGADIR (http://agadir.crg.es/).

Pepsin digestion

Limited proteolysis was carried out using pepsin in 50 mM glycine buffer pH 2.5 and 35 μM of URN1-FF protein. The digestion was performed at 310 K containing an E/S ratio of 1:200 (by weight) and the reactions were quenched after 0.5, 1 and 5 min by adding an appropriate volume of ammonia 2%. The proteolytic mixtures were analyzed by Tricine-SDS/12% PAGE and by MALDI-TOFF MS using an Ultraflex spectrometer (Bruker) operating in linear mode under 20 kV. Samples were prepared by mixing equal volumes of the protein solution and matrix solution (10 mg/ml of sinapic acid dissolved in aqueous 30% acetonitrile with 0.1% TFA) and using the dried droplet method. A mixture of proteins from Bruker (protein calibration standard I; mass range = 3–25 kDa) was used as external mass calibration standard. Peptide fragments were identified by mass fingerprinting analysis.

ACKNOWLEDGMENTS

We acknowledge Dr MJ Macias for providing us the URN1-FF cDNA. This work was supported by BFU2010-14901 from Ministerio de Ciencia e Innovación (MCISpain) and 2009-SGR 760 from AGAUR (Generalitat de Catalunya). V.C. is beneficiary of predoctoral FPU AP 2007-02849 fellowship from the Ministerio de Educación-Spain. S.V. has been granted an ICREA ACADEMIA award.

REFERENCES

1. Dobson CM (2003) Protein folding and misfolding. Nature 426: 884-890.
2. Daggett V, Fersht AR (2009) Protein folding and binding: moving into unchartered territory. Curr Opin Struct Biol 19: 1-2.
3. Chiti F, Dobson CM (2006) Protein misfolding, functional amyloid, and human disease. Annu Rev Biochem 75: 333-366.
4. Nelson R, Eisenberg D (2006) Recent atomic models of amyloid fibril structure. Curr Opin Struct Biol 16: 260-265.
5. Dobson CM (2004) Principles of protein folding, misfolding and aggregation. Semin Cell Dev Biol 15: 3-16.
6. Jahn TR, Radford SE (2005) The Yin and Yang of protein folding. Febs J 272: 5962-5970.
7. de Groot NS, Sabate R, Ventura S (2009) Amyloids in bacterial inclusion bodies. Trends Biochem Sci 34: 408-416.
8. 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.
9. Fandrich M, Fletcher MA, Dobson CM (2001) Amyloid fibrils from muscle myoglobin. Nature 410: 165 - 166.
10. Fandrich M, Forge V, Buder K, Kittler M, Dobson CM, et al. (2003) Myoglobin forms amyloid fibrils by association of unfolded polypeptide segments. Proc Natl Acad Sci U S A 100: 15463-15468.
11. Sirangelo I, Malmo C, Casillo M, Mezzogiorno A, Papa M, et al. (2002) Tryptophanyl substitutions in apomyoglobin determine protein aggregation and amyloid-like fibril formation at physiological pH. J Biol Chem 277: 45887 - 45891.

12. Bedford MT, Leder P (1999) The FF domain: a novel motif that often accompanies WW domains. Trends Biochem Sci 24: 264-265.
13. Gasch A, Wiesner S, Martin-Malpartida P, Ramirez-Espain X, Ruiz L, et al. (2006) The structure of Prp40 FF1 domain and its interaction with the crn-TPR1 motif of Clf1 gives a new insight into the binding mode of FF domains. J Biol Chem 281: 356-364.
14. Allen M, Friedler A, Schon O, Bycroft M (2002) The structure of an FF domain from human HYPA/FBP11. J Mol Biol 323: 411-416.
15. Bonet R, Ramirez-Espain X, Macias MJ (2008) Solution structure of the yeast URN1 splicing factor FF domain: comparative analysis of charge distributions in FF domain structures-FFs and SURPs, two domains with a similar fold. Proteins 73: 1001-1009.
16. Jiang W, Sordella R, Chen GC, Hakre S, Roy AL, et al. (2005) An FF domain-dependent protein interaction mediates a signaling pathway for growth factor-induced gene expression. Mol Cell 17: 23-35.
17. Smith MJ, Kulkarni S, Pawson T (2004) FF domains of CA150 bind transcription and splicing factors through multiple weak interactions. Mol Cell Biol 24: 9274-9285.
18. Jemth P, Gianni S, Day R, Li B, Johnson CM, et al. (2004) Demonstration of a low-energy on-pathway intermediate in a fast-folding protein by kinetics, protein engineering, and simulation. Proc Natl Acad Sci U S A 101: 6450-6455.
19. Jemth P, Day R, Gianni S, Khan F, Allen M, et al. (2005) The structure of the major transition state for folding of an FF domain from experiment and simulation. J Mol Biol 350: 363-378.
20. Jemth P, Johnson CM, Gianni S, Fersht AR (2008) Demonstration by burst-phase analysis of a robust folding intermediate in the FF domain. Protein Eng Des Sel 21: 207-214.
21. Korzhnev DM, Religa TL, Lundstrom P, Fersht AR, Kay LE (2007) The folding pathway of an FF domain: characterization of an on-pathway intermediate state under folding conditions by (15)N, (13)C(alpha) and (13)C-methyl relaxation dispersion and (1)H/(2)H-exchange NMR spectroscopy. J Mol Biol 372: 497-512.

22. Korzhnev DM, Religa TL, Banachewicz W, Fersht AR, Kay LE A transient and low-populated protein-folding intermediate at atomic resolution. Science 329: 1312-1316.
23. Korzhnev DM, Religa TL, Kay LE Transiently populated intermediate functions as a branching point of the FF domain folding pathway. Proc Natl Acad Sci U S A.
24. Korzhnev DM, Vernon RM, Religa TL, Hansen AL, Baker D, et al. Nonnative interactions in the FF domain folding pathway from an atomic resolution structure of a sparsely populated intermediate: an NMR relaxation dispersion study. J Am Chem Soc 133: 10974-10982.
25. Barette J, Velyvis A, Religa TL, Korzhnev DM, Kay LE Cross-Validation of the Structure of a Transiently Formed and Low Populated FF Domain Folding Intermediate Determined by Relaxation Dispersion NMR and CS-Rosetta. J Phys Chem B 116: 6637-6644.
26. Skora L, Becker S, Zweckstetter M (2010) Molten globule precursor states are conformationally correlated to amyloid fibrils of human beta-2-microglobulin. Journal of the American Chemical Society 132: 9223-9225.
27. 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.
28. Ventura S (2005) Sequence determinants of protein aggregation: tools to increase protein solubility. Microb Cell Fact 4: 11.
29. Belli M, Ramazzotti M, Chiti F Prediction of amyloid aggregation in vivo. EMBO Rep 12: 657-663.
30. Castillo V, Grana-Montes R, Sabate R, Ventura S Prediction of the aggregation propensity of proteins from the primary sequence: aggregation properties of proteomes. Biotechnol J 6: 674-685.
31. Bierzynski A, Kim PS, Baldwin RL (1982) A salt bridge stabilizes the helix formed by isolated C-peptide of RNase A. Proc Natl Acad Sci U S A 79: 2470-2474.
32. Shoemaker KR, Kim PS, Brems DN, Marqusee S, York EJ, et al. (1985) Nature of the charged-group effect on the stability of the C-peptide helix. Proc Natl Acad Sci U S A 82: 2349-2353.
33. Hamada D, Kuroda Y, Tanaka T, Goto Y (1995) High helical propensity of the peptide fragments derived from beta-lactoglobulin, a predominantly beta-sheet protein. J Mol Biol 254: 737-746.

34. Jasanoff A, Fersht AR (1994) Quantitative determination of helical propensities from trifluoroethanol titration curves. Biochemistry 33: 2129-2135.
35. Albert JS, Hamilton AD (1995) Stabilization of helical domains in short peptides using hydrophobic interactions. Biochemistry 34: 984-990.
36. Segawa S, Fukuno T, Fujiwara K, Noda Y (1991) Local structures in unfolded lysozyme and correlation with secondary structures in the native conformation: helix-forming or -breaking propensity of peptide segments. Biopolymers 31: 497-509.
37. Nerelius C, Sandegren A, Sargsyan H, Raunak R, Leijonmarck H, et al. (2009) Alpha-helix targeting reduces amyloid-beta peptide toxicity. Proceedings of the National Academy of Sciences of the United States of America 106: 9191-9196.
38. Bartels T, Choi JG, Selkoe DJ (2011) alpha-Synuclein occurs physiologically as a helically folded tetramer that resists aggregation. Nature 477: 107-110.
39. Khurana R, Gillespie JR, Talapatra A, Minert LJ, Ionescu-Zanetti C, et al. (2001) Partially folded intermediates as critical precursors of light chain amyloid fibrils and amorphous aggregates. Biochemistry 40: 3525-3535.
40. Zurdo J, Guijarro JI, Jimenez JL, Saibil HR, Dobson CM (2001) Dependence on solution conditions of aggregation and amyloid formation by an SH3 domain. J Mol Biol 311: 325-340.
41. Jemth P, Day R, Gianni S, Khan F, Allen M, et al. (2005) The structure of the major transition state for folding of an FF domain from experiment and simulation. Journal of Molecular Biology 350: 363-378.
42. Korzhnev DM, Religa TL, Banachewicz W, Fersht AR, Kay LE (2010) A transient and low-populated protein-folding intermediate at atomic resolution. Science 329: 1312-1316.
43. Chiti F, Taddei N, Baroni F, Capanni C, Stefani M, et al. (2002) Kinetic partitioning of protein folding and aggregation. Nat Struct Biol 9: 137-143.
44. Canet D, Last AM, Tito P, Sunde M, Spencer A, et al. (2002) Local cooperativity in the unfolding of an amyloidogenic variant of human lysozyme. Nat Struct Biol 9: 308-315.
45. Niraula TN, Haraoka K, Ando Y, Li H, Yamada H, et al. (2002) Decreased thermodynamic stability as a crucial factor for familial amyloidotic polyneuropathy. Journal of Molecular Biology 320: 333-342.

46. Maurer-Stroh S, Debulpaep M, Kuemmerer N, Lopez de la Paz M, Martins IC, et al. (2010) Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. Nature Methods 7: 237-242.
47. 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.
48. Bryan AW, Jr., Menke M, Cowen LJ, Lindquist SL, Berger B (2009) BETASCAN: probable beta-amyloids identified by pairwise probabilistic analysis. PLoS Computational Biology 5: e1000333.
49. Garbuzynskiy SO, Lobanov MY, Galzitskaya OV (2010) FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. Bioinformatics 26: 326-332.
50. Trovato A, Seno F, Tosatto SC (2007) The PASTA server for protein aggregation prediction. Protein Eng Des Sel 20: 521-523.

Table 1. Thermal denaturation thermodynamic parameters of URN1-FF at different pHs

	T_m (K)		ΔH_m (kcal mol^{-1})	
	CD[1]	Intrinsic fluorescence[2]	CD	Intrinsic fluorescence
pH 2.0	313.5±1.3	310.18±0.3	71.5±1.3	74.2±2.9
pH 2.5	321.5±0.1	318.88±0.1	74.0±1.3	79.8±1.5
pH 3.0	332.8±0.1	335.82±0.06	80.7±1.5	71.8±0.5
pH 5.7	340.9±0.3	341.35±0.07	94.5±6.0	76.5±0.7

[1]Changes in molar ellipticity were monitored at 222 nm.

[2]Changes in intrinsic fluorescence were monitored at at 350 nm.

Table 2. Thermodynamic characterization of URN1-FF at different pHs

	$\Delta G_{f\text{-}u}$[1] (kcal mol^{-1})		$m_{f\text{-}u}$[2] (kcal mol^{-1} M^{-1})		$[U]_{1/2}$[3] (M)	
	CD[4]	Intrinsic fluorescence[5]	CD	Intrinsic fluorescence	CD	Intrinsic fluorescence
pH 2.0	0.36±0.04	0.051±0.08	0.83±0.04	0.43±0.13	0.43	0.19
pH 2.5	1.39±0.13	1.16±0.3	0.98±0.15	0.74±0.5	1.42	1.57
pH 3.0	2.23±0.27	1.77±0.6	1.13±0.07	0.77±0.1	1.97	2.29
pH 5.7	5.33±0.25	5.94±0.18	0.79±0.04	0.88±0.03	6.75	6.75
FBP11[6]	3.62±0.03	3.60±0.02	1.18±0.001	1.14±0.001	3.05	3.16

[1]Gibbs energy of unfolding with urea determined from the equilibrium parameters.

[2]Dependence of the Gibbs energy of unfolding with urea

[3]The urea concentration required to unfold 50% of the protein molecules.

[4]Changes in molar ellipticity were monitored at 222 nm.

[5]Changes in intrinsic fluorescence were monitored at at 350 nm.

[6]The values of the FF domain of FBP11/HYPA were previously reported [18,19].

Table 3. Kinetic parameters for URN1-FF at pH 5.7

	$\Delta G_{f\text{-}u}$[1] (kcal mol^{-1})	$m_{f\text{-}u}$[2] (kcal mol^{-1} M^{-1})	C_m[3] (M)	k_f (s^{-1})	k_u (s^{-1})	m_u (kcal mol^{-1} M^{-1})	m_f (kcal mol^{-1} M^{-1})
298 K	5.57	0.781	7.13	2741±95	0.23±0.11	0.33±0.06	-0.99±0.02
310 K	5.28	0.813	6.50	7022±221	1.33±0.06		
FBP11[4]	3.64	1.16	3.14	2200±90	3.66±0.06	0.154±0.003	1.01±0.02

[1]Gibbs energy of unfolding with urea determined from the kinetic parameters.

[2]Dependence of the Gibbs energy of unfolding with urea.

[3]The urea concentration required to unfold 50% of the protein molecules.

[4]The values of the FF domain of FBP11/HYPA were previously reported [19].

Table S1. Aggregation kinetics of URN1-FF at pH 2.5

	Lag time (min)	k_e (min^{-1})
Thioflavin-T	31	0.0077±0.0004
Scattering	34	0.0102±0.0009

FIGURE LEGENDS

Figure 1. Structure of the URN1-FF domain. At the top, a ribbon representation showing the α-helices in different colors: α-helix 1 in red; α-helix 2 in green, helix 3_{10} in magenta; and α-helix 3 in blue. At the bottom, side chains of Trp 27 and Trp 56 are represented in black and blue, respectively. The N- and C-termini are indicated. The Protein Data Bank accession code for the structure is 2JUC. This figure was prepared with PyMOL.

Figure 2. pH dependence of URN1-FF conformational properties. Protein samples were prepared at 20 μM and were immediately measured by **(a)** far-UV CD, **(b)** tryptophan intrinsic fluorescence and **(c)** ANS fluorescence at 298 K. The fluorescence emission spectrum of ANS in the absence of protein is represented as a dotted line. Soluble URN1-FF species at pH 2.0 (squares), pH 2.5 (diamonds), pH 3.0 (circles), pH 4.0 (crosses) and pH 5.7 (triangles). **(d)** One-dimensional NMR (^{1}H-NMR) spectra of URN1-FF were recorded at 298 K and 600 MHz, using a protein concentration of 35 μM. Two different spectra were collected, at pH 5.7 (above) and pH 2.5 (below).

Figure 3. Thermal and chemical stabilities of URN1-FF at different pHs. Thermal stabilities were studied monitoring the changes by **(a)** far-UV CD at 222 nm and by **(b)** intrinsic fluorescence at 350 nm. Equilibrium urea unfolding curves at 310 K were followed by **(c)** far-UV CD at 222 nm and by **(d)** tryptophan intrinsic fluorescence at 350 nm. Protein samples are represented at pH 2.0 by squares; pH 2.5, diamonds; pH 3.0, circles; and pH 5.7, triangles. All the curves were fitted to a two-state model.

Figure 4. Unfolding and refolding kinetics of URN1-FF. (a) The kinetics of unfolding and refolding for URN1-FF at pH 5.7 were followed by Trp intrinsic fluorescence. Stopped-flow experiments were performed at 298 K (circles) and 310 K (triangles). **(b)** Unfolding rates in 8 M urea at pH ranging from 3.5 to 6.5 were recorded at 310 K. The rate constants were measured under conditions of apparent two-state folding.

Figure 5. Morphological, structural and tinctorial properties of URN1-FF aggregates at different pHs. (a) Representative TEM images of URN1-FF aggregates at 140 μM under different pH conditions incubated at 310 K for one week. From left to right: pH 2.0, pH 2.5, pH 3.0 and pH 5.7. **(b)** Fluorescence emission spectra of thioflavin-T (25 μM) in the absence (dotted line) and in the presence of 10 μM of protein aggregates formed as in (a). **(c)** Absorption spectra of Congo Red (16 μM) in the absence (dotted line) and in the presence of 17 μM of URN1-FF aggregates formed as in (a). The inset shows the difference spectra obtained by subtracting CR-phosphate buffer spectrum from protein in CR-phosphate buffer spectrum. **(d)** Far-UV CD spectra of native URN1-FF (dotted line) and aggregates, using a final concentration of 20 μM. **(e)** Fluorescence emission spectra of ANS (25 μM) collected in the absence (dotted line) and in the presence of protein aggregates (10 μM) formed as in (a). In all the cases, protein aggregates at pH 2.0 are represented as squares; pH 2.5, diamonds; pH 3.0, circles; and pH 5.7, triangles.

Figure 6. URN1-FF aggregation kinetics at pH 2.5. Change of **(a)** ThT fluorescence (25 μM) and **(b)** light scattering at 350 nm during aggregation of URN1-FF at 140 μM, pH 2.5 and 310 K. The insets show some kinetic plots on an expanded time scale.

Figure 7. Prediction of aggregation-prone regions in URN1-FF and peptide design. (a) Aggregation profiles were predicted using Waltz, FOLDAMYLOID and Aggrescan algorithms, from top to bottom. **(b)** At the top, ribbon diagram of the URN1-FF domain showing the designed peptides in different colours: N_t peptide in black, H1 peptide in red, H2 peptide in green, 3_{10} peptide in orange and H3 peptide in blue. At the bottom, amino acid sequences corresponding to each peptide. The residues involved in the formation of α-helices in the native structure are shown in bold.

Figure 8. α-helical structure of synthetic URN1-FF peptides. Synthetic peptides were prepared at 100 μM and pH 2.5. Their far-UV CD spectra were recorded at 298 K in the absence (black) and in the presence of different percentages (v/v) of TFE: 5% (v/v) (blue); 10% (v/v) (orange); 15% (v/v) (green); 20% (v/v) (red); and 25% (v/v) (yellow). **(a)** N_t, **(b)** H1, **(c)** H2, **(d)** 3_{10} and **(e)** H3 peptides. **(f)** Plot of the molar ellipticity at 222 nm versus the TFE concentration for all the synthetic peptides: N_t (green); H1 (black); H2 (blue); 3_{10} (orange) and H3 (red) peptides. Each value

represents the difference between the molar ellipticity at different concentrations of TFE and the molar ellipticity in the absence of TFE.

Figure 9. Aggregation properties of synthetic URN1-FF peptides. URN1-FF peptides were prepared at 1500 μM for the N_t peptide and 600 μM for the rest of the peptides, pH 2.5, 310 K in the absence and in the presence of TFE. Protein samples incubated for one week were diluted to a final concentration of 100 μM and their far-UV CD spectra (solid lines) were compared with those of peptides under non-aggregating conditions (dashed lines). **(a)** N_t, **(b)** H1, **(c)** H2, **(d)** 3_{10} and **(e)** H3 peptides in 0% (black), 15% (green) and 25% (red) (v/v) TFE. **(f)** Fluorescence emission spectra of ThT (25 μM) in the absence (black dashed lines) and in the presence of 40 μM aggregated H1 (red) and H3 (blue) peptides. In the inset, we show aggregated N_t (green), H2 (orange) and 3_{10} (grey) peptides. In all cases, samples in the absence of TFE are represented as solid lines, and samples with 25% (v/v) TFE are indicated as dashed lines. TEM images of synthetic peptides are shown in the absence of TFE: **(g)** N_t, **(h)** H1, **(i)** H2, **(j)** 3_{10} and **(k)** H3. The insets show the aggregated peptides in the presence of 25% (v/v) TFE.

Figure 10. Limited proteolysis of URN1-FF at low pH. Pepsin digestion was carried out at 310 K in 50 mM glycine, pH 2.5, and 35 μM URN1-FF, using an E/S ratio of 1:200 (by weight). **(a)** Time course proteolysis monitored by Tricine-SDS/12% (w/v) PAGE gel with Comassie blue staining. The star indicates aprotinin (6.5 kDa). **(b)** MALDI-TOFF MS before 0.5 min and 5 min after pepsin digestion. The peptides observed with a statistically significant change in abundance are denoted with their molecular weight. The star indicates a 3560 Da peak, corresponding to half URN1-FF mass. **(c)** Amino acid sequence of URN1-FF domain. The arrows indicate the pepsin cleavages inside the H1 (residue 16) and H3 (residue 52) segments. Residues forming α-helices are shown in bold. Secondary structure elements are represented above and below the sequence.

Figure 1

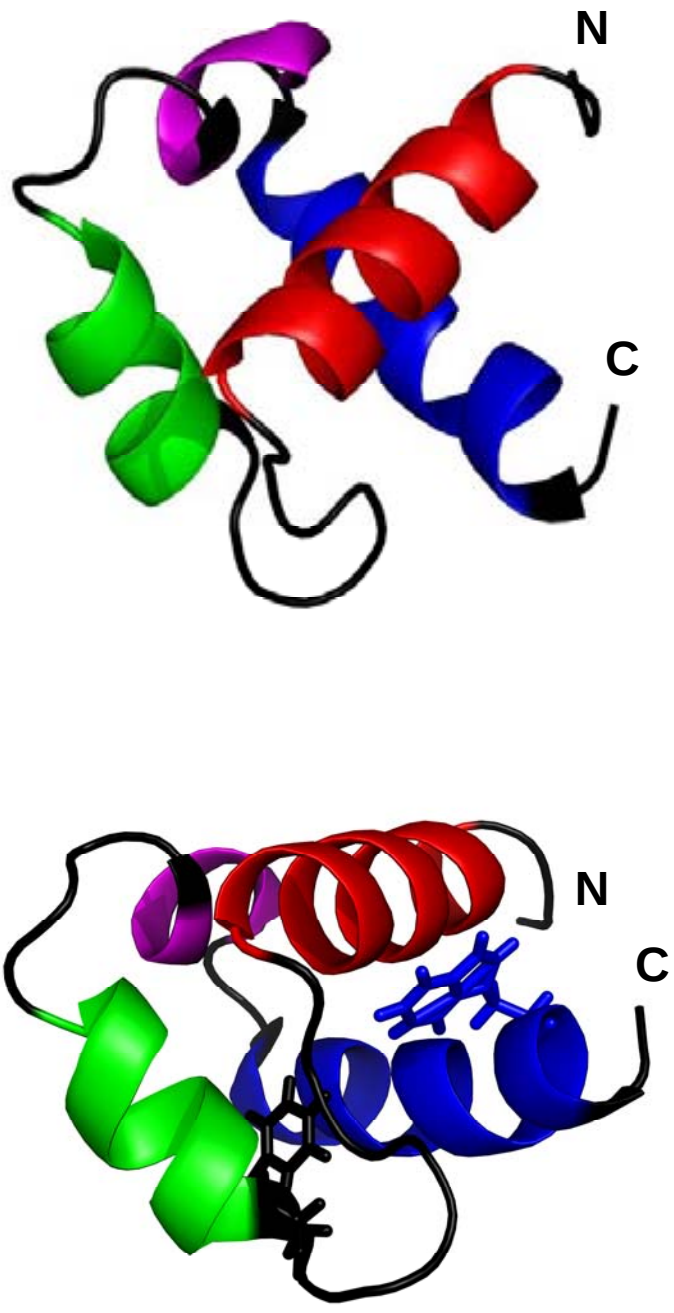


Figure 2

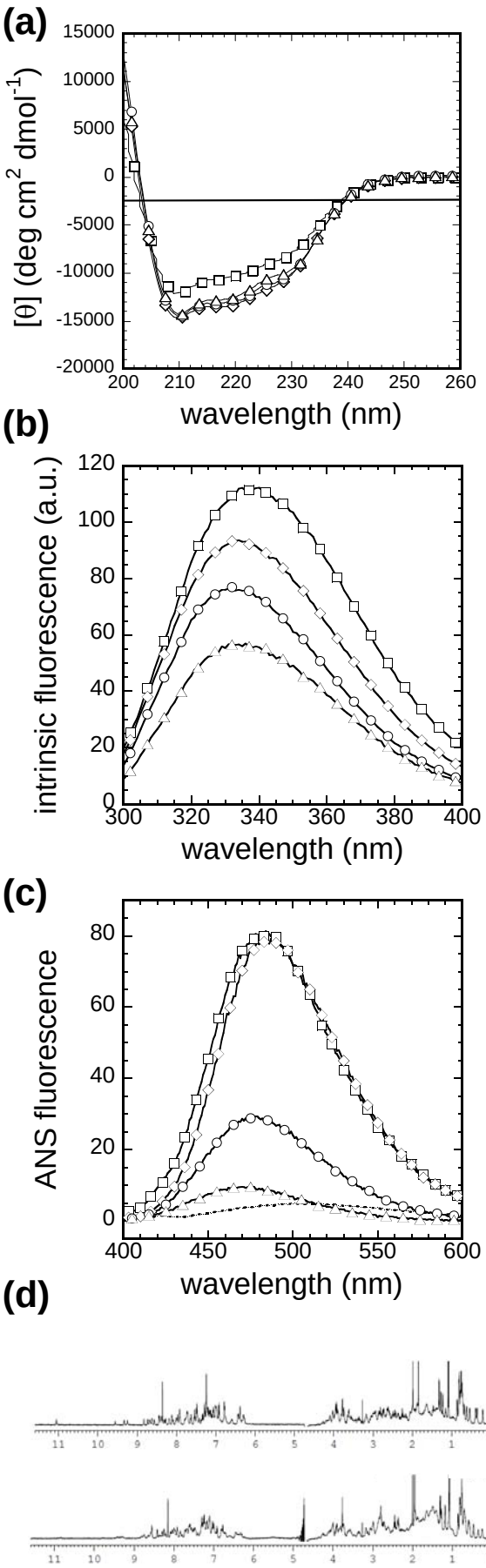

(a)
15000
10000
5000
0
-5000
-10000
-15000
-20000
200
210
220
230
240
250
260
[θ] (deg cm2 dmol-1)
wavelength (nm)
(b)
120
100
80
60
40
20
0
300
320
340
360
380
400
intrinsic fluorescence (a.u.)
wavelength (nm)
(c)
80
60
40
20
0
400
450
500
550
600
ANS fluorescence
wavelength (nm)
(d)
11
10
9
8
7
6
5
4
3
2
1
0
ppm

Figure 3

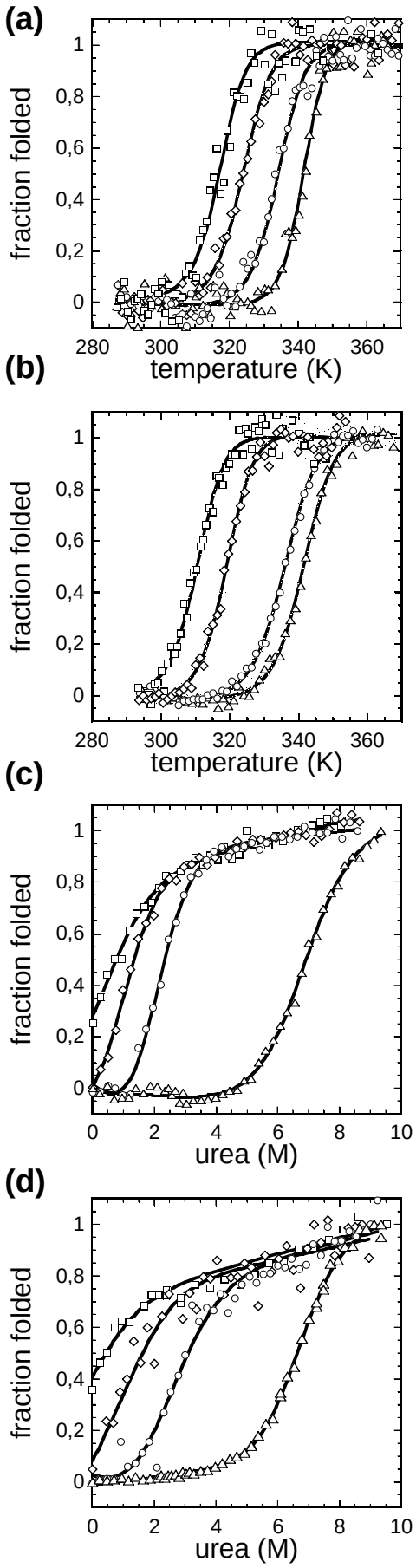

(a)
fraction folded
1
0,8
0,6
0,4
0,2
0
280
300
320
340
360
temperature (K)
(b)
fraction folded
1
0,8
0,6
0,4
0,2
0
280
300
320
340
360
temperature (K)
(c)
fraction folded
1
0,8
0,6
0,4
0,2
0
0
2
4
6
8
10
urea (M)
(d)
fraction folded
1
0,8
0,6
0,4
0,2
0
0
2
4
6
8
10
urea (M)

Figure 4

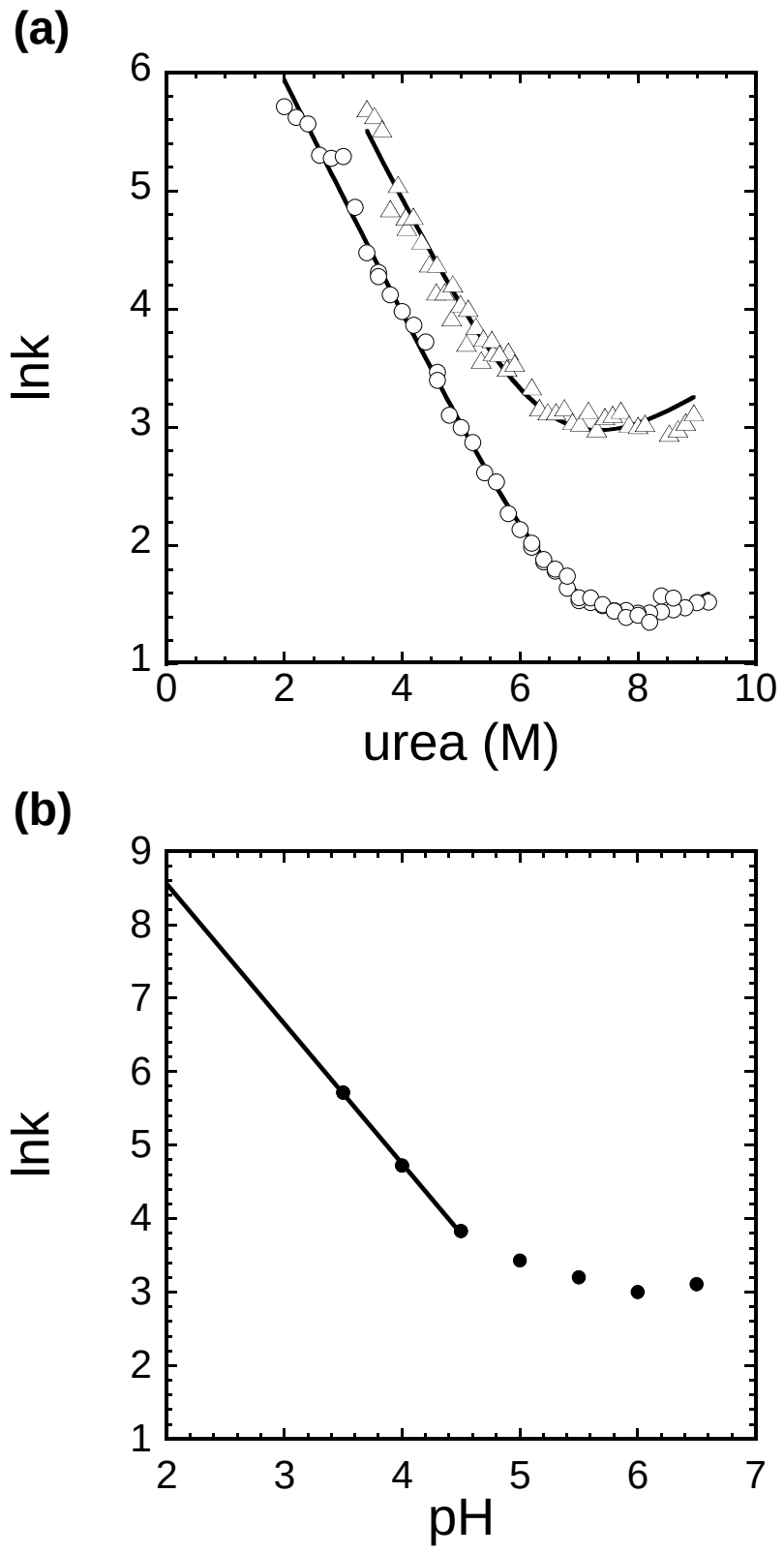


Figure 5

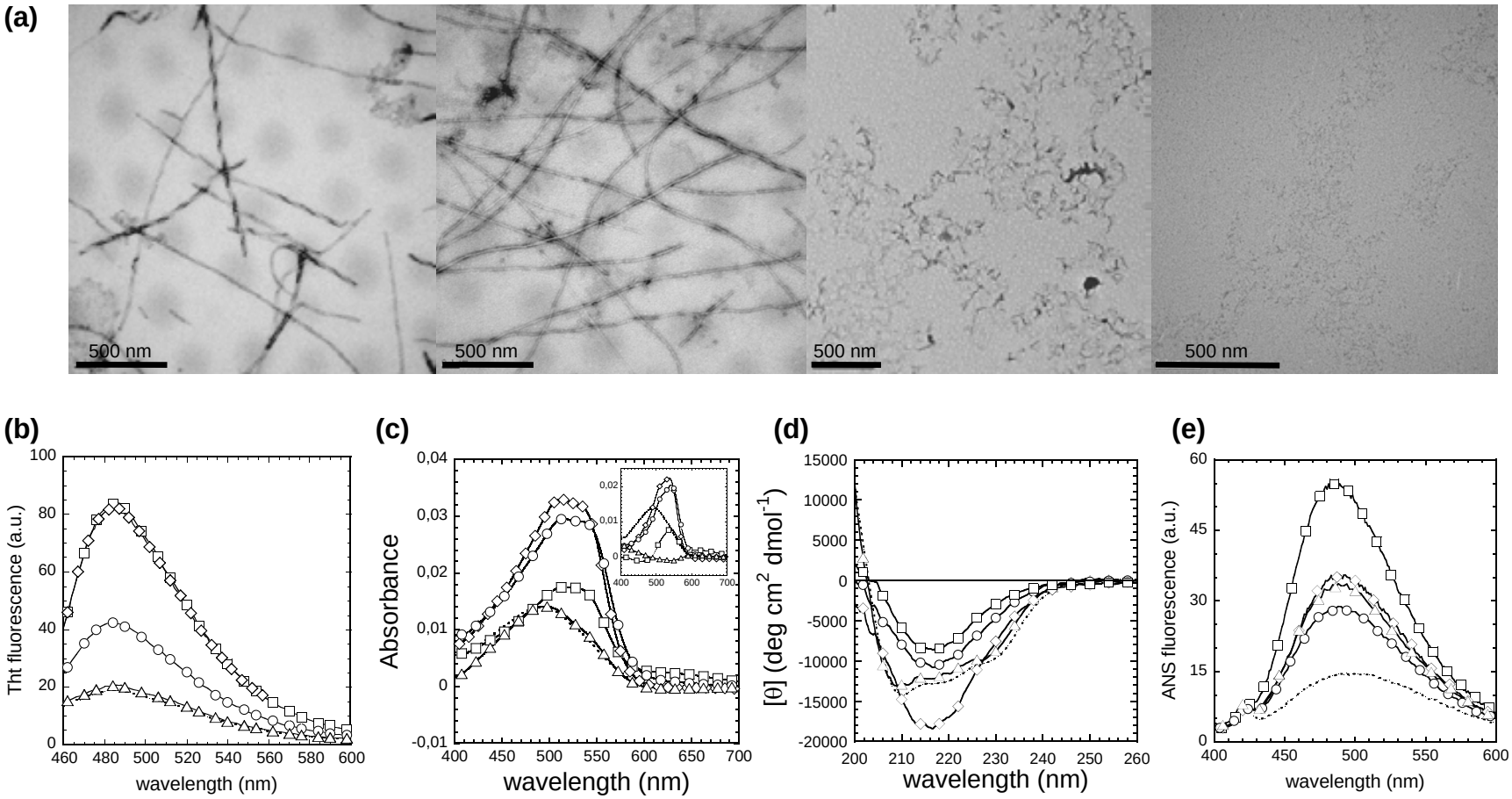

(a)
500 nm
500 nm
500 nm
500 nm
(b)
Tht fluorescence (a.u.)
wavelength (nm)
(c)
Absorbance
wavelength (nm)
(d)
[θ] (deg cm^2 dmol^{-1})
wavelength (nm)
(e)
ANS fluorescence (a.u.)
wavelength (nm)

Figure 6

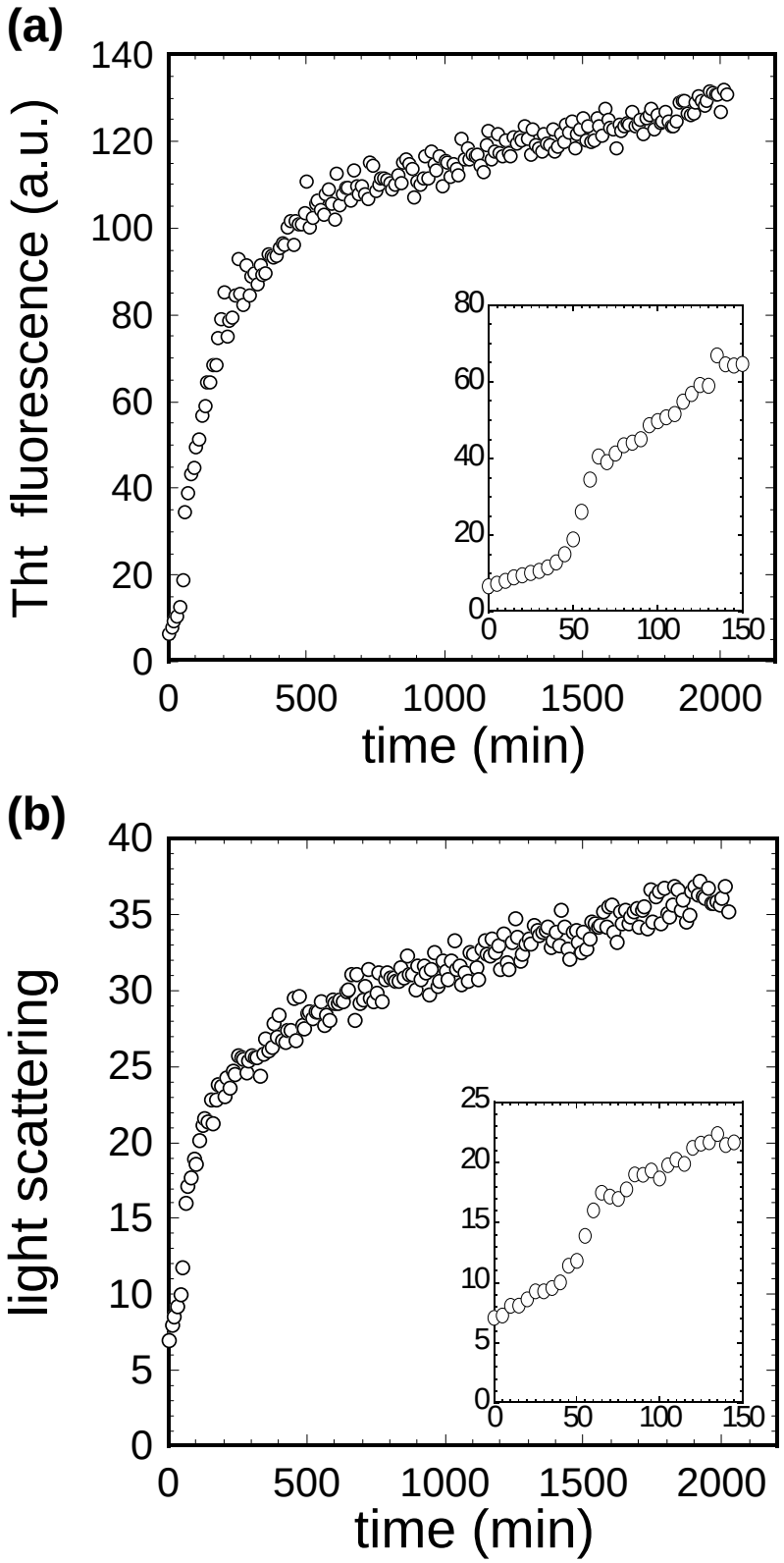
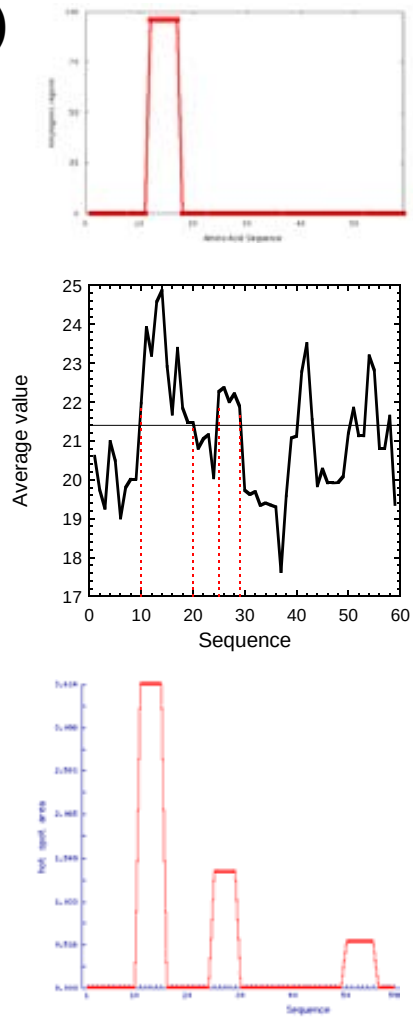
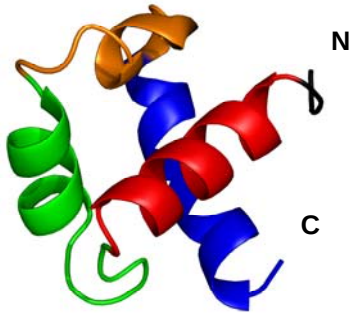

(a)
Tht fluorescence (a.u.)
0
20
40
60
80
100
120
140
0
500
1000
1500
2000
time (min)
0
20
40
60
80
0
50
100
150
(b)
light scattering
0
5
10
15
20
25
30
35
40
0
500
1000
1500
2000
time (min)
0
5
10
15
20
25
0
50
100
150

Figure 7

(a)



(b)



Peptide	Sequence	Residues
N_t	GAMGDID	1-7
H1	**ERNIFFEL FDRY**K	8-20
H2	LDKFSTW **SLQSKKIE**	21-35
3_{10}	NDPD**FYKI** RD	36-45
H3	**DTVRESLF EEWC**GE	46-59

Figure 8

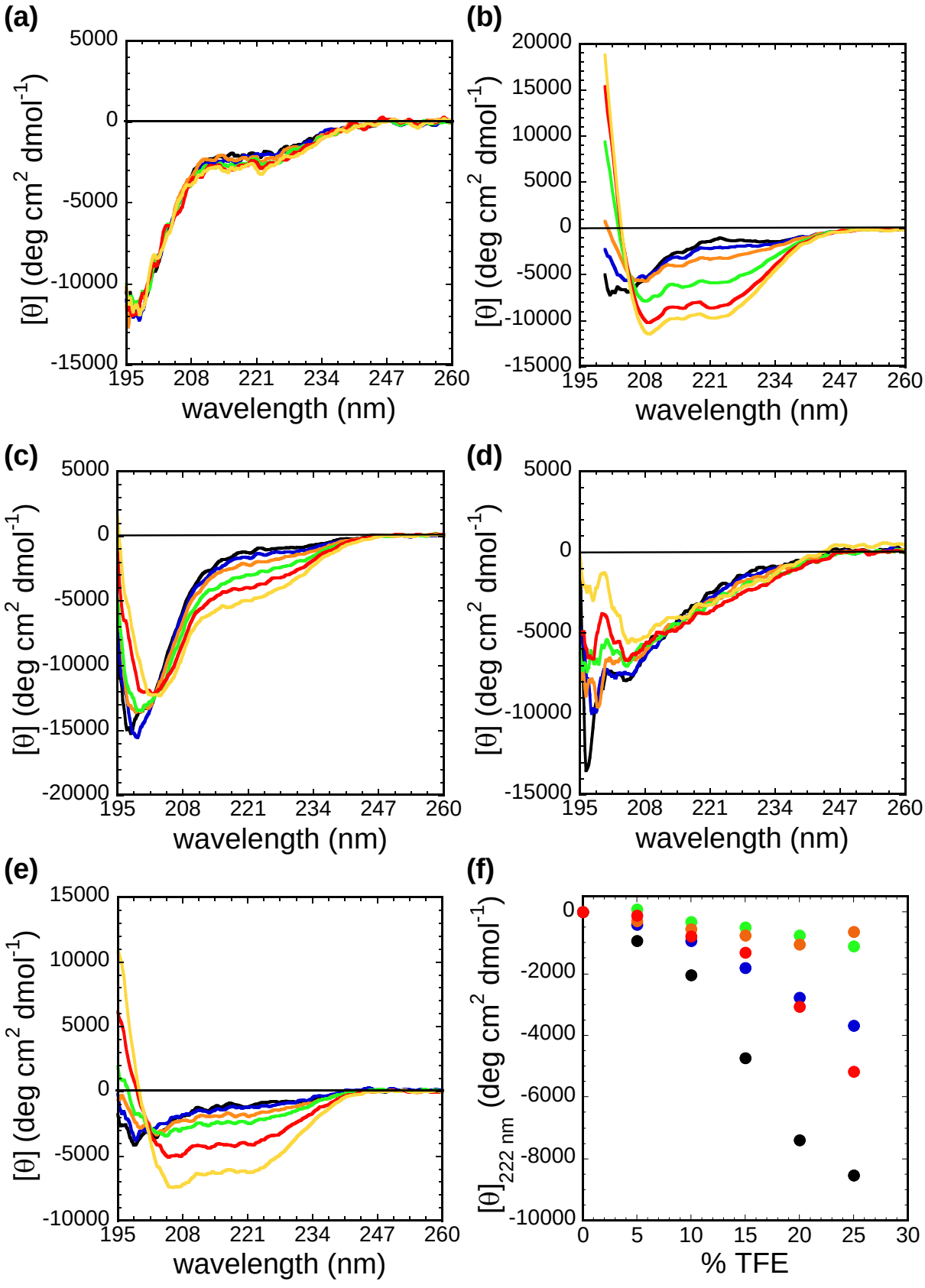

(a)
(b)
(c)
(d)
(e)
(f)
[θ] (deg cm2 dmol-1)
wavelength (nm)
[θ]222 nm (deg cm2 dmol-1)
% TFE
195
208
221
234
247
260

Figure 9

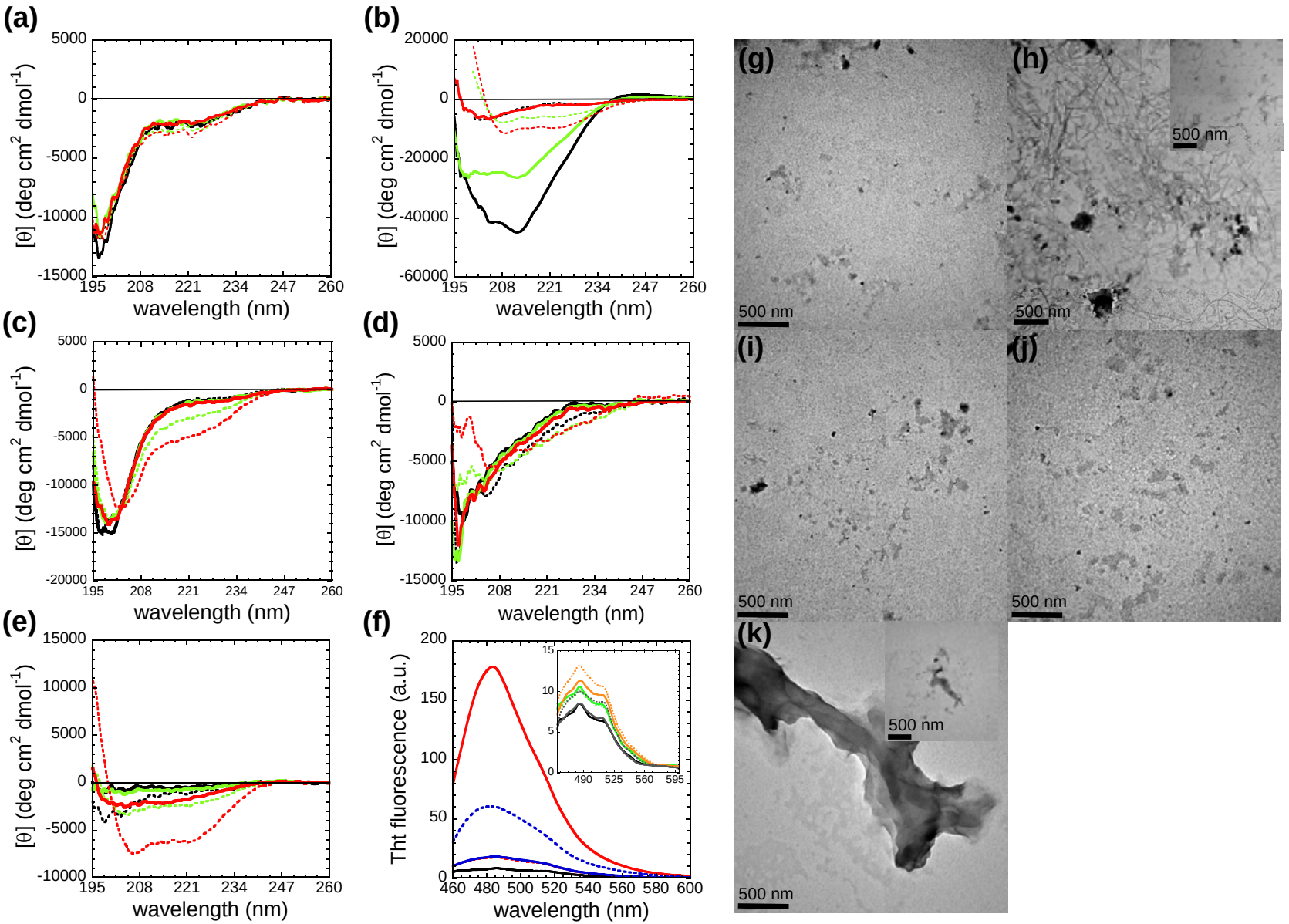

(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
(k)
[θ] (deg cm2 dmol-1)
wavelength (nm)
Tht fluorescence (a.u.)
500 nm

Figure 10

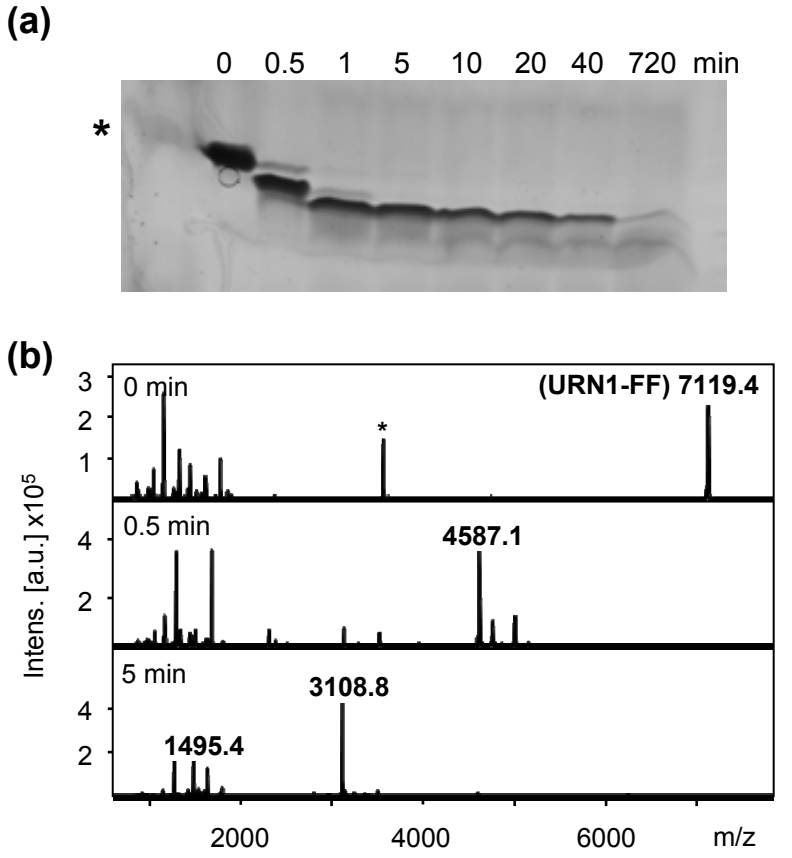

(a)
0 0.5 1 5 10 20 40 720 min
*
(b)
0 min
(URN1-FF) 7119.4
*
0.5 min
4587.1
5 min
3108.8
1495.4
Intens. [a.u.] x10^5
2000
4000
6000
m/z
(c)
10
20
30
GAMGDIDERNIFFELFDRYKLDKFSTWSLQSKKIE
40
50
NDPDFYKIRDDTVRESLFEEWCGE

Figure S1

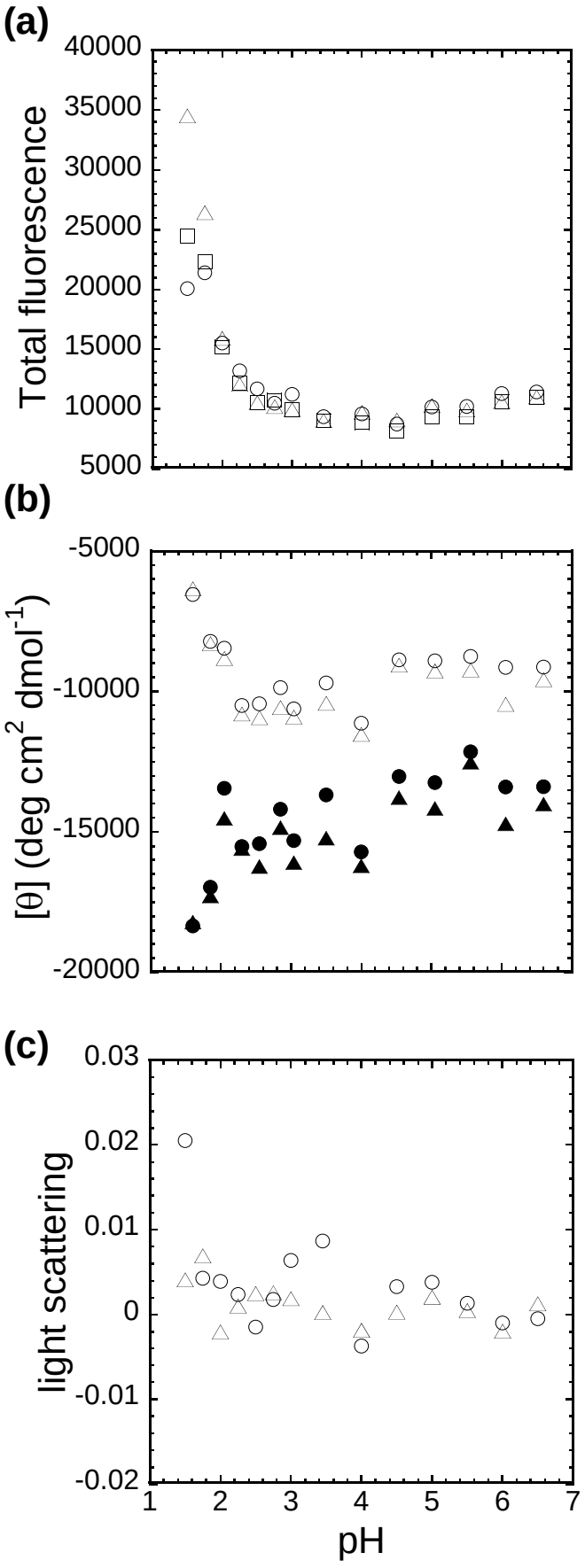
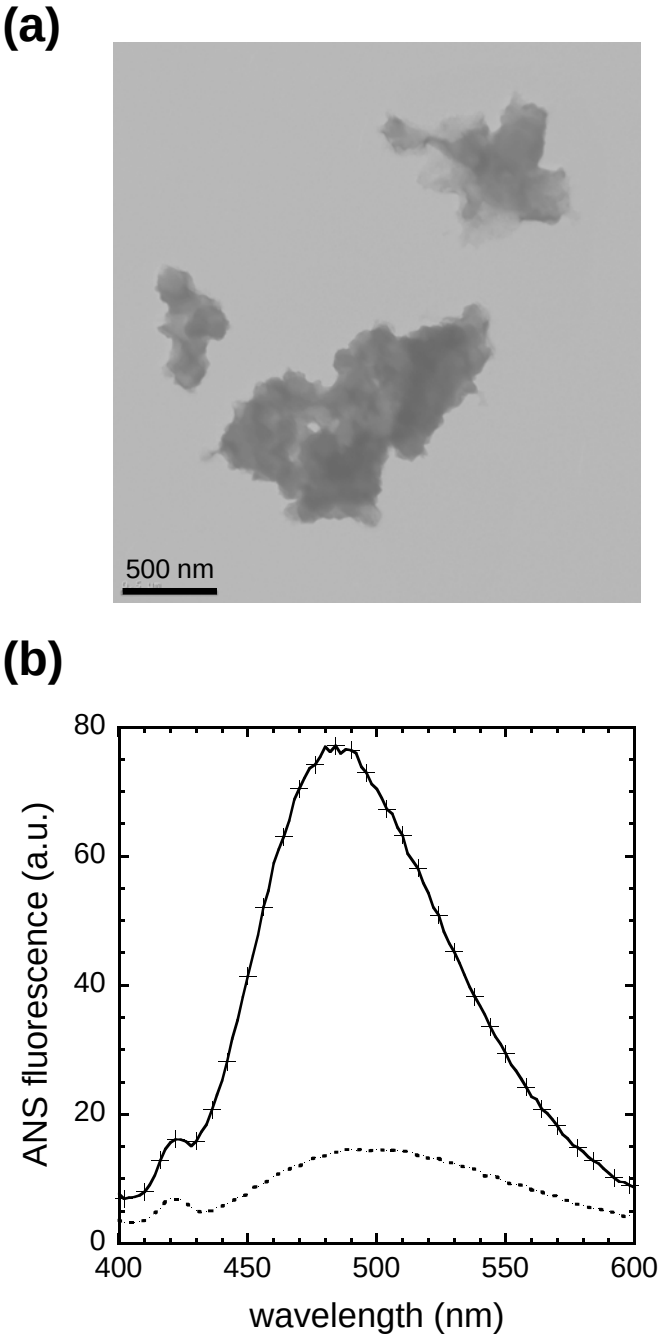

(a)
Total fluorescence
40000
35000
30000
25000
20000
15000
10000
5000
(b)
[θ] (deg cm2 dmol-1)
-5000
-10000
-15000
-20000
(c)
light scattering
0.03
0.02
0.01
0
-0.01
-0.02
1
2
3
4
5
6
7
pH

Figure S2


(a)
500 nm
(b)
ANS fluorescence (a.u.)
0
20
40
60
80
400
450
500
550
600
wavelength (nm)

SUPPLEMENTARY MATERIAL

Figure S1. Evolution of the conformational properties of soluble URN1-FF species with time. Protein samples were prepared at low protein concentration (20 μM) and at pH ranging from 1.5 to 6.5. **(a)** Total tryptophan intrinsic fluorescence was measured after 1.5 h (triangles), 6 h (squares) and 24 h (circles) of sample preparation. **(b)** Far-UV CD signals at 230 nm (empty symbols) and 215 nm (filled symbols) were recorded after 3 h (triangles) and 24 h (circles) of protein dissolution. **(c)** Light scattering was followed at 350 nm after 3 h (triangles) and 24 h (circles) of protein preparation.

Figure S2. Conformational properties of URN1-FF aggregates at pH 4.0. (a) Representative TEM image of URN1-FF aggregate at 140 μM, pH 4.0 incubated at 310 K for one week. **(b)** Fluorescence emission spectra of ANS (25 μM) collected in the absence (dotted line) and presence of 10 μM of protein aggregates (crosses).