Universitat Autònoma de Barcelona

Departament de Bioquímica i Biologia Molecular
and
Institut de Biotecnologia i Biomedicina

Analysis of Different Evolutionary Strategies to Prevent Protein Aggregation

Ricardo Graña Montes

Bellaterra, December 2014

Universitat Autònoma de Barcelona

Departament de Bioquímica i Biologia Molecular
and
Institut de Biotecnologia i Biomedicina

Analysis of Different Evolutionary Strategies to Prevent Protein Aggregation

Doctoral thesis submitted by Ricardo Graña Montes in candidacy for the degree of Ph.D. in Biochemistry, Molecular Biology and Biomedicine from the Universitat Autònoma de Barcelona.

The work described herein has been performed at the Department de Bioquímica i Biologia Molecular and at the Institut de Biotecnologia i Biomedicina, under the supervision of Prof. Salvador Ventura Zamora.

Ricardo Graña Montes Prof. Salvador Ventura Zamora

Bellaterra, December 2014

SUMMARY IN ENGLISH

In the last 15 years, the study of protein aggregation has evolved from a mostly neglected topic of protein chemistry to a highly dynamic research area which has expanded its implications through different fields including biochemistry, biotechnology, nanotechnology and biomedicine. The analysis of protein aggregation has attracted a particular interest in the biomedical and biotechnological areas. Because, on one side, the formation of insoluble protein deposits is associated to an increasing number of human disorders, many of which present fatal pathological consequences. And on the other hand, aggregation is a frequent shortcoming in the recombinant expression of proteins at the industrial level, such as in the production of proteinaceous therapeutic agents like antibodies. Consequently, the survey of mechanisms to prevent protein aggregation is currently the focus of deep investigation with the aim to develop preventive or therapeutic methods for the intervention of these depositional disorders and to enhance the yield in the biotechnological production of proteins.

The power of the computational tools developed to predict protein aggregation has fostered the identification of the determinants influencing the aggregation of polypeptides and has allowed to investigate how the selective pressure to avoid aggregation has shaped the cellular proteomes along evolution. From these analyses, different mechanisms to prevent protein aggregation have emerged ranging from negative design strategies found in polypeptide sequences and structures, to the characterization of the factors governing the cellular machinery in charge of the protein quality control.

The present thesis provides a multiperspective analysis for the detailed characterization of several of these mechanisms evolved to confront the risk of protein aggregation. In this sense, the use of a variety of specific proteic models of aggregation or different sets of proteins with related properties has allowed to analyze particular strategies to avoid aggregation in depth. At the same time, the approach based on the study of closely related ensembles of proteins has allowed to identify functional constraints that limit the evolutionary selection against aggregation. More specifically, the work presented here addresses the effect of restricting the configurational freedom of the polypeptide chain by disulfide cross-linking on the aggregation process, as well as the impact over this phenomenon exerted by the presence of intrinsically disordered protein regions. Additionally, the regulation of cellular protein abundance as a function of protein aggregation propensity has also been surveyed. On the other hand the analysis centered on the study of closely related proteins has revealed how the requirements to fold efficiently and to maintain the catalytic activity constrain the minimization of the aggregation propensity of proteins. These analyses highlight particularly the interplay between folding and aggregation, in such a way that the analysis of the aggregational properties of polypeptides allows to forecast the mechanism of folding of certain kind of proteins.

RESUM EN CATALÀ

En els darrers 15 anys, l'estudi de l'agregació de proteïnes ha evolucionat de ser una part de la química de proteïnes que tradicionalment generava poc interès, a convertir-se en una àrea d'investigació dinàmica que ha ampliat el seu abast a diferents camps de recerca incloent-hi la bioquímica, la biotecnologia, la nanotecnologia y la biomedicina. Dins d'aquests camps, l'anàlisi de l'agregació de proteïnes un interès particularment rellevant en les àrees biomèdica i biotecnològica. Això es degut, per una banda, a que la aparició de depòsits proteics insolubles está relacionada amb un nombre creixent de malalties humanes, moltes d'elles amb conseqüències fatals per als malalts. Per l'altre costat, l'agregació proteica es una complicació habitual en la expresió recombinant de proteïnes a nivell industrial, com ara a la producció d'agents terapèutics de naturalessa proteica, com els anticossos. Conseqüentment, l'exploració de mecanismes que permetin prevenir l'agregació proteica es subjecte actualment d'una intensa tasca d'investigació adreçada a desenvolupar métodes per a la prevenció i l'intervenció terapèutiques d'aquestes malalties greus; i també amb l'objectiu d'incrementar els rendiments en la producció biotechnològica de proteïnes.

La gran capacitat de les eines computacionals desenvolupades per tal de predir l'agregació proteica ha donat un gran impuls als esforços destinats a identificar els determinants de l'agregació de polipèptids, i també ha permès investigar com la pressió selectiva contra l'agregació ha moldeat els proteomes cel·lulars al llarg de l'evolució. A partir d’aquests anàlisis, s’han pogut identificar diferents mecanismes evolucionats per prevenir l’agregació proteica que van des de estratègies de disseny negatiu que s’han detectat en seqüències i estructures de proteïnes, fins a la caracterització dels factors que governen la maquinària cel•lular encarregada del control de qualitat proteic.

En aquesta tesi es proporciona un anàlisis des de diferents perspectives per assolir una caracterització detallada de diversos d’aquests mecanismes que han sorgit al llarg de l’evolució per fer front al risc d’agregació. Amb aquest objectiu, s’han fet servir tant models proteics d’agregació específics com diferents conjunts de proteïnes amb propietats relacionades, que han permès analitzar en profunditat estratègies contra l’agregació. Al mateix temps, l’aproximació basada en l’estudi de conjunts de proteïnes íntimament relacionats també ha permès identificar limitacions funcionals que limiten la selecció a nivell evolutiu contra l’agregació. De manera més específica, el treball que es presenta aquí aborda l’efecte de la restricció de la llibertat conformacional de la cadena polipeptídica, causada per l’establiment d’un pont disulfur, sobre el procés d’agregació; també s’ha analitzat l’impacte de regions intrínsicament desestructurades en aquest mateix fenomen d’agregació. A més també s’ha investigat la regulació de l’abundància proteica dins la cèl·lula en funció de la tendència a agregar específica de les proteïnes. Per altra banda, l’anàlisi centrat en l’estudi de proteïnes relacionades ha revelat com els requeriments per assolir un plegament eficient i per mantenir l’activitat catalítica limiten la disminució de la càrrega d’agregació de les proteïnes. Aquests

anàlisis posen especialment en relleu el balanç entre plegament funcional i agregació, de manera que la caracterització de les propietats aggregatives de determinats polipèptids permet predir el seus mecanismes de plegament.

LIST OF PUBLICATIONS

This thesis is composed of the following published works:

I. **Ricardo Graña-Montes**, Virginia Castillo, &. Salvador Ventura: The Aggregation Properties of *Escherichia coli* Proteins is Associated with their Cellular Abundance. *Biotechnology Journal* 6(6) pp. 752–760 (2011)

II. **Ricardo Graña-Montes**, Natalia Sánchez de Groot, Virginia Castillo, Javier Sancho, Adrian Velázquez-Campoy & Salvador Ventura: Contribution of Disulfide Bonds to Stability, Folding, and Amyloid Fibril Formation: the PI3-SH3 Domain Case. *Antioxidants & Redox Signaling* 16(1) pp. 1-15 (2012)

III. **Ricardo Graña-Montes**, Ricardo Sant'Anna de Oliveira & Salvador Ventura: Protein Aggregation Profile of the Human Kinome. *Frontiers in Physiology* 3 – 438, (2012)

IV. **Ricardo Graña-Montes**, Patrizia Marinelli, David Reverter & Salvador Ventura: N-terminal Protein Tails Act as Aggregation Protective Entropic Bristles: the SUMO Case. *Biomacromolecules* 15 (4) pp 1194-1203, (2014)

V. **Ricardo Graña-Montes**, Hugo Fraga, Ricard Illa, Giovanni Covaleda & Salvador Ventura:Association between Foldability and Aggregation Propensity in Small Disulfide-Rich Proteins. *Antioxidants & Redox Signaling*, 21(3) pp 368-383, (2014)

Other articles co-authored which are not part of this thesis:

VI. Virginia Castillo, **Ricardo Graña-Montes**, Raimon Sabaté &. Salvador Ventura: Prediction of the Aggregation Propensity of Proteins from the Primary Sequence: Aggregation Properties of Proteomes. *Biotechnology Journal* 6(6) pp. 674-685 (2011) *review article*

VII. Raimon Sabaté, Alba Espargaró, **Ricardo Graña-Montes**, David Reverter & Salvador Ventura: Native Structure Protects SUMO Proteins from Aggregation into Amyloid Fibrils. *Biomacromolecules* 13(6) pp. 1916-1926 (2012)

VIII. **Ricardo Graña-Montes**, & Salvador Ventura: About Targets and Causes in Protein Folding. *Journal of Biomolecular Structure and Dynamics* 31(9) pp. 970-972 (2013)

IX. Ricardo Sant'Anna, Carolina Braga, Nathalia Varejão, Karinne M. Pimenta, **Ricardo Graña-Montes**, Aline Alves, Juliana Cortines, Yraima Cordeiro, Salvador Ventura & Debora Foguel: The Importance of a Gatekeeper Residue on the Aggregation of Transthyretin: Implications to Transthyretin-related Amyloidoses. *Journal of Biological Chemistry* 289(41) pp. 28324-28337, (2014)

X. Javier Garcia-Pardo, **Ricardo Graña-Montes**, Marc Fernàndez-Mendez, Àngels Ruyra, Nerea Roher, Francesc X. Aviles, Julia Lorenzo & Salvador Ventura: Amyloid Formation by Human Carboxypeptidase D Transthyretin-like Domain under Physiological Conditions. *Journal of Biological Chemistry* -published online- (2014)

CONTENTS

LIST OF ABBREVIATIONS

APR	Aggregation-prone Region
AFM	Atomic Force Microscopy
ALS	Amyotrophic Lateral Sclerosis
ANS	4,4′-bis(1-anilinonaphthalene 8-sulfonate)
ATP	Adenosine triphosphate
bis-ANS	1-anilinonaphthalene 8-sulfonate
BPTI	Bovine Pancreatic Trypsin Inhibitor
CD	Circular Dichroism
CO	Contact Order
CR	Congo Red
Cter	carboxyl terminus
DRDs	Disulfide-rich Domains
EGF	Epidermal growth factor
FTIR	Fourier-transform Infrared Spectroscopy
GFP	Green Fluorescent Protein
GTP	Guanosine triphosphate
IDPRs	Intrinsically Disordered Protein Regions
IDPs	Intrinsically Disordered Proteins
LA5	Ligand binding module 5 of the Low-density Lipoprotein
LCI	Leech Carboxypeptidase Inhibitor
LDTI	Leech-derived Tryptase Inhibitor
LRO	Long-range Order
mRNA	messenger Ribonucleic Acid
NCC	Nucleated Conformational Conversion mechanism
NMR	Nuclear Magnetic Resonance
Nter	amino terminus
NvCI	*Nerita versicolor* Inhibitor
PDB	Protein Data Bank
PI3-SH3	Src Homology 3 domain of Phosphatidylinositol 3-kinase
PQC	Protein Quality Control machinery
SIMs	SUMO Interacting Motif
ssNMR	solid state Nuclear Magnetic Resonance
SUMO	Small Ubiquitin-like MOdifier
TAP	Tick Anticoagulant Peptide
TEM	Transmission Electron Microscopy
Th-T	Thioflavin-T
TS	Transition State
TSE	Transition State Ensemble
Ubl	Ubiquitin-like
UPS	Ubiquitin-proteasome System

CHAPTER 1.- INTRODUCTION

1.1.- Protein Folding

Proteins, or polypeptides, are biological polymers, which are made, in general, from a combination of 20 different amino acids (Creighton 1992). They are the main executors of cellular processes, and in order to exert their biological functions most of them require to adopt a specific three-dimensional structure, also known as the native state (Dobson 2003b). Protein folding accounts for the process by which a polypeptide chain departing from mostly unstructured conformations (commonly referred to as the unfolded or the denatured state) attains its native structure. By the early 60's, several works by Christian B. Anfinsen and colleagues, focusing on the refolding of ribonuclease A, allowed them to conclude that the folding of a protein is primarily determined by its amino acid sequence, and subsequently drove them to establish the "thermodynamic hypothesis" for protein folding or **"Anfinsen's postulate"**, which states that the three-dimensional structure of a protein in its native, physiological environment corresponds to the global minimum of its Gibbs energy function (Anfinsen 1973). We nowadays know, however, that the native state does not necessarily constitute the global minimum of the Gibbs function for a certain polypeptide, since more stable states exist that might be reached under some circumstances, even under physiological conditions (Eichner & Radford 2011; Knowles et al. 2014). Among those states we find different types of protein aggregates, which will be discussed below.

1.1.a.- Evolution of Protein Folding Paradigms

On a foundational paper, Cyrus Levinthal postulated in 1968 that if proteins do not face any other constrain to fold than reaching the energetic minimum, then a polypeptide chain departing from the unfolded state would be expected to experience random conformational fluctuations until reaching that minimum. Levinthal showed that in case the folding process was to proceed through a random search, it would require an astronomic time to be completed, a consideration which was in sharp contrast with the sub-minute folding reactions already experimentally characterized at that time (Levinthal 1968; Levinthal 1969). The apparent paradox revealed by Levinthal, lead him and other researchers to propose that proteins cannot actually explore the whole conformational space but they should fold following specific pathways. Perhaps, at that time, the concept of pathway was more influenced by the knowledge about metabolic pathways (Kim & Baldwin 1982; Kim & Baldwin 1990) and that led many researchers in the field to hypothesize folding pathways were somewhat a limited succession of states with a defined conformation between the unfolded protein and the native state, so the efforts of the folding community moved towards the attempt to characterize such intermediate states. Folding intermediates were detected, analyzed, and even isolated for certain proteins undergoing cis/trans Pro isomerization (Garel & Baldwin 1973; Garel et al. 1976) or disulfide bonding (Creighton 1977; Creighton 1978), processes we nowadays know to constitute particular cases, rather than generic mechanisms of folding. . The reversible folding of proteins had already been proposed as a transition between two principal, unfolded and native, states (Lumry & Biltonen 1966), and a number of proteins were known to follow single exponential unfolding kinetics representative of two-states folding, though they were considered as rare specimens in accordance with the dominant intermediate-mediated folding paradigm.

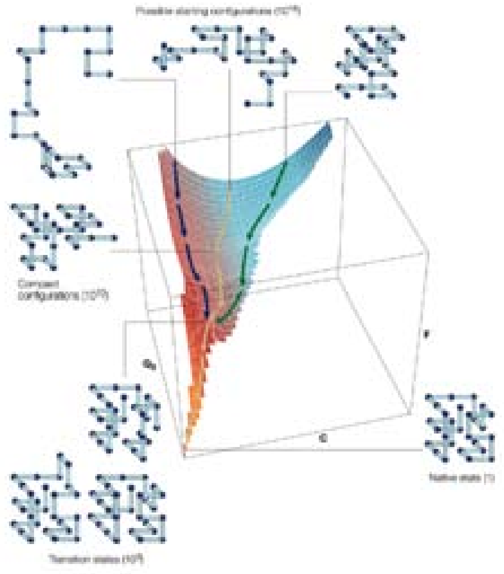

Possible starting configurations (10^{16})
Compact configurations (10^{10})
F
C
Native state (1)
Transition states (10^{3})

Figure 1.- Representation of the Energy Landscape for a Putative Protein. Energy landscape derived from the computational simulation of a simplified representation of a small protein employing a “bead and string” model. The energy of the system decreases as the configurational freedom of the polypeptide chain becomes more restricted. Reproduced form (Dobson 2003a) with permission.

In the early 90’s kinetic studies started to gain popularity, and the analysis of protein folding and unfolding constants in the light of the transition state (TS) theory (Matthews 1987) became common. More precisely, several works by Alan R. Fersht and co-workers settled up a new experimental framework to analyze protein folding, including the establishment of a series of conditions to determine whether a protein folded or not according to a two state regime (Jackson & Fersht 1991). The application of concepts from the organic physical-chemistry field in order characterize transition states in the protein folding reaction (Matouschek & Fersht 1993); and most notably, the advances in cloning technology allowed the generalization of protein engineering analysis by amino acid substitutions and the introduction of the Φ-analysis method to derive structural insights about the nature of the TS (Matouschek et al. 1989; Itzhaki et al. 1995). Therefore, the efforts in the field shifted to the characterization of the metastable species occurring along folding, including -but not merely focusing on- intermediates. Together with the advances in the experimental arena, extremely valuable insights came from the theoretical side after the introduction of statistical mechanics models based on polymer physics (Bryngelson & Wolynes 1987; Leopold et al. 1992). These models relied on simplified representations of polypeptides like, for instance, that of beads linked by strings in a cubic lattice -beads representing amino acids with different properties and strings the peptide bond between them- (Leopold et al. 1992). Computer simulations employing those models provided increasing levels of detail by showing protein folding as a diffusive process in which the Gibbs energy of the system decreases as the degrees of freedom of the chain become more and more constrained (Leopold et al. 1992). The interpretation of such simulations brought a new perspective to the field, sometimes referred to as the **"New View"**, in which the folding of proteins is regarded as a series of transitions between microscopic states possessing more or less related configurations (Onuchic et al. 1997). These models also led to the popularization of the funnel-like diagrams, which have witnessed a huge success due the intuitiveness they allow for the interpretation of folding pathways; however they have been frequently misinterpreted (Karplus 2011) as the energy landscape for folding themselves, when they are merely low-dimensionality representations of certain components of the energy function, typically the proteins internal energy versus a couple of parameters describing different configurational degrees of freedom of the polypeptide (Figure 1).

1.1.b.- Mechanisms of Protein Folding

Together with the interest in dissecting the pathways to attain correct protein folding, researchers also became interested in defining the mechanism describing such reactions (Figure 2).

A first model was proposed based on a simple topological analysis of different protein structures, which was later termed the **"nucleation-growth"** mechanism. It argued that protein structures could be subdivided in "regions", defined as portions of the structure that could be enclosed within a compact volume. In order to overcome the "Levinthal´s paradox", the folding of a "region" proceeded through the formation of an initial nucleus of several residues forming native structure, whose growth would take place by the addition of neighboring peptide stretches to the nucleus so the structure was progressively expanded until reaching the final native conformation (Wetlaufer 1973). This model implied a progressive gain of native structure, which was judged incompatible with the prevailing paradigm at that time of intermediate-dominated pathways, so it attracted little attention.

One different proposal which was also introduced in the 70's, the **"diffusion-collision"** or **"framework"** model, stated that the folding of proteins proceeded in a hierarchical manner, with portions of the polypeptide sequence or "microdomains", usually associated to secondary structure elements, which can sample the conformational space in an independent manner. These elements can diffuse and eventually collide and dock together to adopt the correct tertiary structure (Ptitsyn & Rashin 1975; Karplus & Weaver 1976; Karplus & Weaver 1994).

A third mechanism, denominated **"hydrophobic collapse"**, was postulated based on earlier studies in the 50's and 60's highlighting the relevance of interactions between hydrophobic residues in the stabilization of protein structures (Kauzmann 1959; Tanford 1962). According to this model, the first stage in the folding of a polypeptide consists in the collapse of the polypeptide chain by the establishment of unspecific interactions involving hydrophobic residues, excluding them from solvent water molecules. Further reorganization of both the polypeptide backbone and the amino acid sidechains within this collapsed conformation would trigger the development of secondary and tertiary structure, leading to consolidation of the native structure (Dill et al. 1995). Criticism to this model arose by the observation that such a hydrophobic collapse would lead to an excess of unspecific non-native interactions that would hamper chain reorganization. A subsequent development of this model introduced the concept of **molten globule** to explain the mechanism of folding (Ptitsyn 1995) arguing the polypeptide chain acquires secondary structure concomitantly with its collapse.

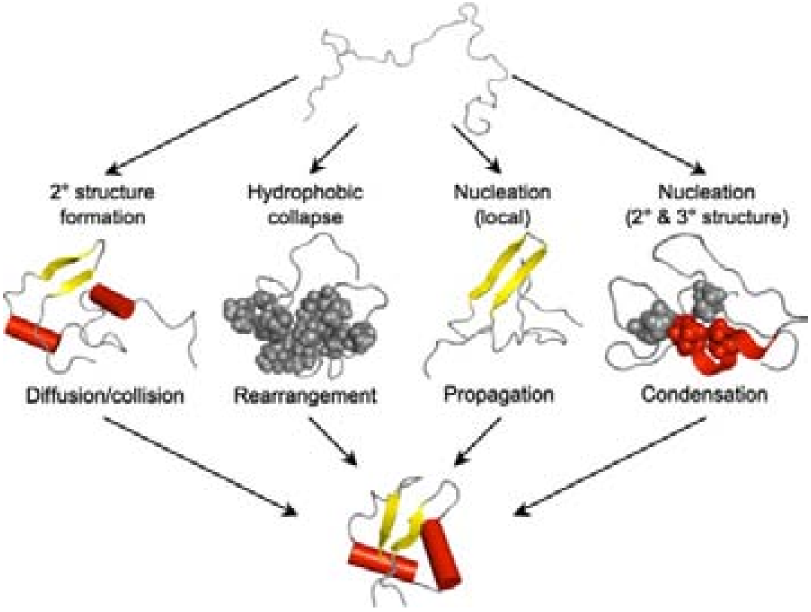


Figure 2.- Schematic Representation of the Classical Mechanisms of Protein Folding. The general features of acquisition of secondary and tertiary structure are depicted for each one of the classical mechanisms proposed for the folding of proteins. Reproduced from (Nickson & Clarke 2010).

Since both the hydrophobic-collapse and the framework models include the existence of detectable intermediates along the folding pathway, they readily became the predominant mechanisms to rationalize protein folding reactions.

Nonetheless, the thorough Φ-analysis of the transition state of CI2, the first model to have been demonstrated to follow a pure two-state folding and unfolding kinetics (Jackson & Fersht 1991), led to a further development of the nucleation model, **"nucleation-condensation"** mechanism. Following this model, the collapse of the polypeptide chain is driven by the formation of a folding nucleus, which characterizes the transition state, and is achieved by the establishment of long-range interactions that, at the same time, stabilize secondary structure elements formed concomitantly to polypeptide chain condensation (Itzhaki et al. 1995).

This work paved the way to further analysis of the transition state of proteins adopting different tertiary structures. The accumulation of kinetic data from different protein models, coupled to the analysis of the residue contact network in a given fold, revealed rather surprisingly how a simplistic measurement of the topology of the polypeptide chain allows to rationalize its folding kinetics (Baker 2000). Protein topology can be defined by simple parameters reflecting the distance between interacting residues in the polypeptide structure, such as the Contact Order (CO) (Plaxco et al. 1998) or the Long Range Order (LRO) parameters (Gromiha & Selvaraj 2001), and their computed values yield a good correlation with experimental folding contacts. This provides a plain, yet elegant framework to interpret the folding mechanism of a polypeptide. In fact, the relevance of the protein topology on its folding has been highlighted by computational simulations constrained with experimental Φ-values (Lindorff-Larsen et al. 2004; Lindorff-Larsen et al. 2005), which show that acquisition of the native topology is the major determinant for the formation of the transition state ensemble, and the contact network established by a few key residues suffices to lead the attainment of such topology.

1.1.c.- Current View of Protein Folding

Nowadays, protein folding is regarded as a chain-like diffusion process where the polypeptide gains compactness as it reaches its native structure. This process is characterized by a series of states, that correspond to local minima of the energy function associated to the polypeptide conformation in an aqueous solvent, and whose relative stability is sufficient for them to be significantly populated in spite of the thermal fluctuations of the system.

The interconversion between these states, which is constrained by its kinetic accessibility, defines the folding pathway of a specific polypeptide. Such states shall not be regarded as static conformations with precisely defined structures, but rather as ensembles of intimately related conformations, with similar energies, that can easily interconvert between them. In the simplest case, only two states can be populated, namely the unfolded and the native state ensembles, and the transition between them usually involves the crossing of an energetic barrier that characterizes the transition state ensemble (TSE). Intermediate state ensembles are defined when conformations other than those corresponding to the unfolded and the native states are significantly populated. Such intermediates may be classified as: (i) on-pathway, when they allow for the progression of the folding reaction towards the native state, or (ii) off-pathway, when they constitute dead-ends that cannot reach the native conformation.

Regarding the mechanisms of protein folding, a unified model has been proposed based on the observation that homologous proteins with the same fold may follow either the nucleation-condensation or the framework models depending on the intrinsic propensity of their sequences to adopt the native secondary structure (Gianni et al. 2003; Banachewicz et al. 2011).

The current unified view of protein folding mechanisms consist of a model sliding between the framework and the nucleation-condensation mechanisms depending on the secondary structure propensity of the protein and the relative strength of its folding nucleus (Nickson et al. 2013). For a polypeptide with strong secondary structure propensity, secondary structure elements may readily form before the establishment of contacts stabilizing the tertiary structure, so the protein folds via a framework mechanism. When the tendency to adopt secondary structure is weaker and a stable folding nucleus exists, the acquisition of the native topology may concomitantly stabilize secondary structure formation in a typical nucleation-condensation model. In the case where no strong secondary structure propensity is present, and neither there is a sufficiently stable nucleus, the folding of the polypeptide may be guided by the diffusion and collision of previously formed, marginally stable, secondary structure elements following a framework-like mechanism.

1.2.- Folding of Disulfide-rich Proteins

Proteins containing disulfide bonds have always played a crucial role in the development of the field of protein folding, from the very first experiments performed by Anfinsen and co-workers on the renaturation of RNase A -which are considered as the foundation of this research area- to the efforts directed towards the isolation of folding intermediates during the 70’s and 80’s. However, it was later realized that the folding of proteins enriched in disulfide bonds constitutes a particular case within protein folding due to the strong impact disulfide formation has on the kinetics and thermodynamics of the folding process (Arolas et al. 2006).

First, the covalent bond between cysteine pairs provides a great stabilization of the native state, which is mainly attributed to the reduction of the configurational entropy of the unfolded state that arises from intrachain (Poland & Scheraga 1965; Lin et al. 1984).

This increased stability of the native conformation in disulfide-rich proteins provides them with a higher strength towards structural damage caused by, ie. oxidative species or proteolytic enzymes. This property is particularly relevant for those proteins performing their function in harsh environments, such as those inherent to the extracellular space (Zavodszky et al. 2001; Arolas & Ventura 2011). Accordingly, a larger half-life has been reported for these proteins (REF). Although the major role of disulfide bonding is considered to be related to the stability and folding of the polypeptides, the oxidative formation and reductive breakage of disulfide linkage has also been shown to regulate the function of certain proteins (Hogg 2003).

The folding of polypeptides enriched in disulfide bonds is generically named **oxidative folding** and comprises two phenomena: the formation of the native cysteine pairings or disulfide

regeneration and the acquisition of the native structure or **conformational folding** (Wedemeyer et al. 2000; Narayan et al. 2000). The native disulfide pairing corresponds to that the protein exhibits once it has attained its final folded state.

The formation of a covalent bond between two cysteine residues is a second order reaction, which depends on the local intramolecular concentration of their reactive thiol groups; the latter being defined by the spatial proximity between the cysteines. The process of disulfide formation also depends on the thiol reactivity of the specific cysteine residue, and is influenced by the dielectric constant of its environment (Wedemeyer et al. 2000), which may be modulated depending on the nature of the neighboring side chains. Another extremely relevant variable influencing the covalent linkage of cysteines is the accessibility of the reactive groups; this is particularly important for the exchange that can be established between free thiol groups and already formed disulfides during oxidative folding or reshuffling. Disulfide bonds buried within the structure hamper any further exchange (Wedemeyer et al. 2000; Narayan et al. 2000).

As mentioned above, disulfide-rich proteins populate intermediates due to the marginal stabilization provided by the successive formation of disulfide bonds. Covalent bonding between cysteines allows for both native and non-native cysteine pairings to be formed along the folding pathways. Intermediates allowing the progression towards the native state (productive) usually protect their formed disulfides through burial into the structure, while those representing dead-ends in the course of folding (non-productive) tend to be “disulfide-insecure” with disulfide bonds being exposed to favour reshuffling by exchange with free thiols (Narayan et al. 2000).

According to the properties of the intermediates populated during the oxidative folding reaction of a disulfide-rich polypeptide, its folding mechanism is defined by the degree of homogeneity of its intermediates, the nature of the cysteine pairing these intermediates possess -either native or non-native-, and whether fully oxidized isomers with non-native disulfides (scrambled isomers) are populated or not (Arolas et al. 2006).

Based on this definition, two extreme mechanisms have been described, which are exemplified by the resolved folding pathways of bovine pancreatic trypsin inhibitor (BPTI) and hirudin, respectively (Figure 3). BPTI folds through a discrete number of intermediates possessing native disulfide connectivity, which also adopt native-like conformations (Weissman & Kim 1991; Weissman & Kim 1992); in this kind of mechanism, the intermediates formed by cysteine crosslinking appear to guide the conformational folding towards the native state. In contrast, hirudin folds through the population of a highly heterogeneous ensemble of intermediates with no prevalence of intermediate species with native disulfide connectivity. Moreover, fully oxidized scrambled isomers are also populated, which experience disulfide reshuffling to yield the native conformation (Chatrenet & Chang 1992; Chatrenet & Chang 1993). Two major phases may be differentiated in the oxidative folding of hirudin-like proteins. First, the formation of intermediates with unspecific connectivity is associated to a packing stage resulting in a substantial decrease in the configurational entropy of the polypeptide chain, without a specific gain in native-like structure. A consolidation step follows in which reshuffling between scrambled isomers straitens the conformational search for the native state.

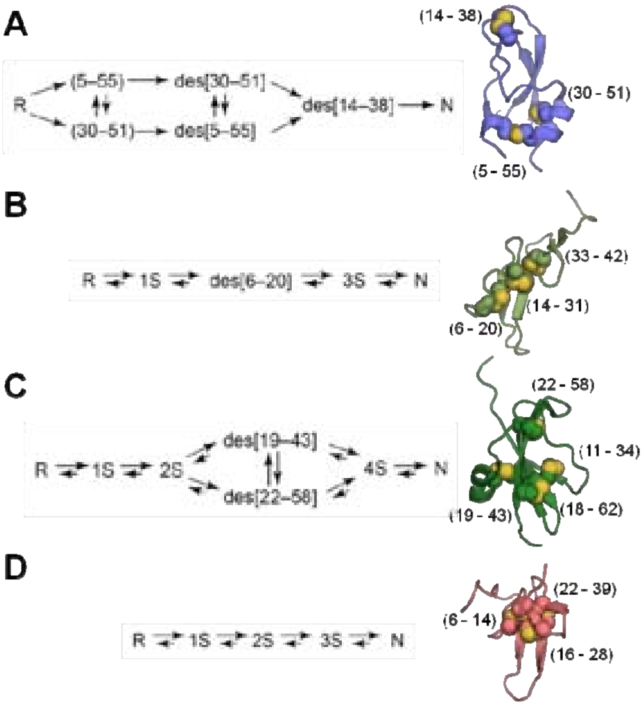


Figure 3.- Oxidative Folding Pathways Resolved Experimentally. The oxidative folding pathways of A) BPTI (PDB 1pit), B) EGF (PDB 1ivo), C) LCI (PDB 1dtv) and D) hirudin (PDB 1hic) are shown; the R and N symbols represent the fully reduced and the fully oxidized and native conformations, respectively, S indicates intermediates with variable conformations and disulfide connectivities but the same number of formed disulfide, and the "des" prefix denotes and intermediate with native-like conformation and disulfide connectivity which only lacks one disulfide bond. Disulfide bonds in the native conformation are shown as yellow spheres. Adapted from (Arolas et al. 2006).

Intermediate folding mechanisms between these two extremes have been reported with mixed characteristics from the BPTI and hirudin-like models. This is exemplified by the folding mechanisms of the epidermal growth factor (EGF) and the leech carboxypeptidase inhibitor (LCI) whose oxidative folding proceeds first by the formation of a heterogeneous ensemble of intermediates which readily lead later to defined intermediates with two or three native disulfide bonds for EGF (Wu et al. 1998; Chang et al. 2001) and LCI (Salamanca et al. 2003; Arolas et al. 2004), respectively. However, these native-like intermediates do not yield the native conformation directly but act rather as kinetic traps requiring reshuffling of their native disulfides to produce scrambled isomers; these scrambled forms do later rearrange to attain the final folded conformation.

As described above, the chemistry of cysteine residues involves a competition between oxidation, reshuffling and reduction reactions during oxidative folding. These processes do not take place at random inside the cell but are subject of precise regulation. Since the cytoplasmic environment usually presents a reductive nature it is not surprising that proteins enriched in disulfide bonds are rarely found in this compartment. Disulfide bond formation takes place in the endoplasmatic reticulum (ER) and the intermembrane mitochondrial space (Fraga & Ventura 2013) of eukaryotic cells or in the periplasm of bacteria (Mamathambika & Bardwell 2008). Moreover, the oxidation of cysteine pairs to form disulfide bonds is catalyzed in vivo by the action of specific enzymes, such as those belonging to the protein disulfide isomerase (PDI) and the oxidoreductase disulfide bond protein (Dsb) families in eukaryotes and bacteria, respectively (Frand et al. 2000; Denoncin & Collet 2013), which promote disulfide exchange, thus catalyzing the conversion of scrambled isomers into the native conformation.

1.3.- Intrinsically Disordered Proteins

For many years the work on protein biophysics has been dominated by the structure-to-function paradigm, which considers that a protein sequence encodes a specific three-dimensional structure that allows for the development of one or several functions (Wright & Dyson 1999). However, an increasing number of proteins lacking a defined equilibrium conformation in its native state are being discovered (Tompa 2012). In addition, the analysis of the structures deposited in the Protein Data Bank has revealed that a majority of proteins harbor continuous regions with missing electron density (Le Gall et al. 2007), a feature commonly attributed to the high flexibility of these stretches. These “unfolded” or “unstructured” polypeptides and protein regions are attracting increasing scientific attention. Although many terms have been coined to categorize this kind of proteins and regions, “intrinsic disorder” is considered the key feature defining their conformational properties while still accounting for their structural diversity, nowadays the denominations Intrinsically Disordered Proteins (IDPs) and Intrinsically Disordered Protein Regions (IDPRs) are preferred (Uversky 2013a). A first extensive analysis of the IDPs biochemical properties revealed these proteins possess a high net charge and low hydrophobicity (Uversky et al. 2000). Later, compositional analysis showed that IDPs are particularly depleted in aromatic (Trp, Phe, Tyr) and aliphatic amino acids (Ile, Leu, Val), as well as in Asn and Cys -relative to the ensemble of structured globular proteins- while they are enriched in polar and charged residues (Lys, Arg, Ser, Gln, Glu) and Pro (Radivojac et al. 2007). The persistent bias presented by IDPs in their primary structure has allowed to develop a multitude of bioinformatic tools to predict structural disorder from sequence (Ferron et al. 2006; He et al. 2009; Jin & Liu 2013) that have permitted to evaluate the IPD content in the proteomes from a wide variety of organisms. Proteome-wide analysis of intrinsic disorder has revealed that 10-35% and 15-45% of proteins in prokaryotes and eukaryotes, respectively, are predicted to possess large disordered regions of at least 30 residues (Tompa 2012). Interestingly, when considering the average content of disorder in proteins, a gap is observed between prokaryotes and eukaryotes, with the first possessing an average disorder below 27% and the latter above 32% (Uversky 2013a). In spite of the criteria selected to account for protein disorder, a tendency is clear for the more complex eukaryotic cells to be more enriched in IDPs and IDPRs. These findings have been linked to the functional roles associated to proteins presenting extensive disordered regions, being predominantly involved in signaling and regulation (Tompa 2012). A survey of the functional annotations correlated with predicted protein disorder revealed, indeed, that these kind of proteins are involved in key cellular processes such as transcription and translation, different phases of the cell cycle, differentiation, and cell death, among other relevant processes (Xie et al. 2007). IDPs have also been implicated in functions of cellular recognition such as those performed by RNA and protein chaperones (Tompa & Csermely 2004). Altogether, the functions described for IPDs rely to a great extent in a specific but transient binding to their partners in the development of such function, typically proteins or nucleic acids. It has been argued that the conformational plasticity inherent to IDPs and IDPRs provides functional advantages for this specific type of binding, such as the ability to display multiple interaction sites -which can even overlap between them-, thus allowing IDPs to be involved in different modes of interaction, like binding of a single IDPR to different structural patterns or interaction of multiple proteins with a single IDP. These binding attributes are consistent with the molecular functions described for IDPRs in signaling interaction networks (Dunker et al. 2005), as well as their action as scaffolds, facilitating the spatial and temporary coordination in the interaction of other proteins (Cortese et al. 2008).

Also, the extended nature of this kind of proteins confers them with a greater capture radius and a higher interaction surface per residue and accessibility, which allows fostering the establishment of interactions. In addition, their flexibility is considered to favour their functional regulation via proteolytic degradation (Uversky 2013a). IDPs and IDPRs properties have been postulated to allow for a specific binding with low strength of interaction, a modality that provides fast association but also rapid dissociation rates, which are considered relevant aspects in regulatory processes where rapid activation is as important as rapid shutdown.

Although IDPs are characterized by their inability to spontaneously fold into a defined three-dimensional conformation, it is known that they can undergo structural gain upon binding to their partners (Spolar & Record 1994; Dyson & Wright 2002) or due to environmental changes (Uversky 2009). However, disorder-to-order transition is not a strict requirement for IDP and IDPR function and substantial disorder may be retained even after binding to their partners, in what have been termed “fuzzy complexes” (Tompa & Fuxreiter 2008). The description of IPDs properties, with their particular attributes, may lead to the impression that there exists a bimodal distribution between completely ordered/structured and absolutely disordered/unstructured proteins. A more realistic view presents a continuum of different degrees of disorder content within protein structures; according to Vladimir N. Uversky, between rigid **ordered structures devoid of IDPRs** and **completely unstructured IDPs** a variety of structures with different proportion of disorder may be found including **ordered proteins with confined intrinsically disordered segments**, typically at their amino or carboxyl terminus or in linker regions, **structured polypeptides with a significant content of IDPRs, molten globule-like collapsed IDPs** and **pre-molten globule-like extended IDPs**.

1.4.- Protein Aggregation

As it has been described before, the folding of a polypeptide chain into its native structure is a complex process, thus errors along folding may occur and drive proteins to incorrectly folded or misfolded states, which may possess still certain stability in the physiological environment and, therefore, start to accumulate (Figure 4). The inability of a protein to achieve or maintain its native conformation, resulting in the formation of different types of aggregates, is associated to an increasing number of human pathologies (Table 1) ranging from neurodegenerative disorders, such as Alzheimer's and Parkinson's diseases, to different amyloidoses, characterized by the formation of large proteinaceous deposits, diabetes mellitus type II or even certain types of cancers (Selkoe 2003; Chiti & Dobson 2006; Invernizzi et al. 2012). This fact, together with the implications on the biotechnological production of proteins, has pushed the study of protein aggregation to evolve from a barely neglected area of protein chemistry to a highly dynamic research field nowadays.

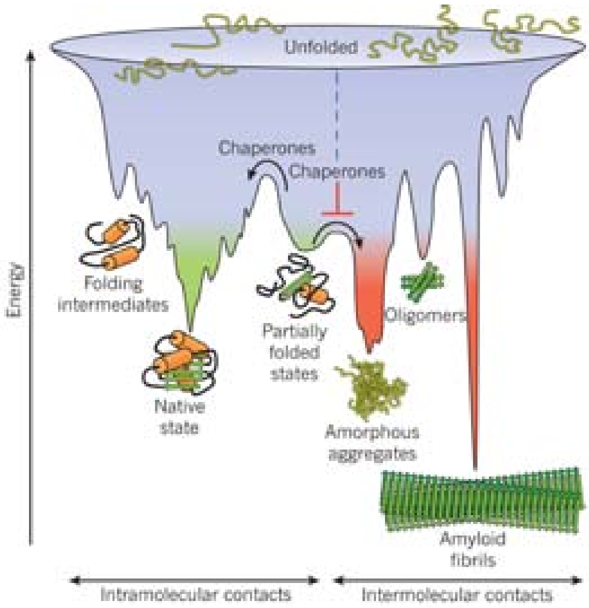


Figure 4.- Schematic Representation of the Hypothetical Energy Landscape Illustrating the Competition between Folding and Aggregation. The folding of a polypeptide chain to its functional native state is guided by the establishment of specific intramolecular contacts; however, the formation of intermolecular interactions may lead to the population of non-functional ordered structures with a higher relative stability. The access to such conformations is usually prevented by the action of molecular chaperones assisting the correct folding to the native structure or by blocking the formation of intermolecular contacts. Reproduced form (Hartl et al. 2011) with permission.

1.4.a.- Historical Overview

The first references we have regarding the presence in human tissues of anomalous deposits that might be associated to amyloid-like protein aggregates, come from 17th century autopsy reports (Kyle 2001). Later 19th century medical reports tell us about unusual inclusions in different organs, mainly in the liver and the spleen, of pale-colored and dense substances, which possibly also corresponded to accumulations of amyloid proteins (Doyle 1988). In 1854, Rudolph Virchow was the first who employed the term "amyloid" to name cerebral structures with abnormal appearance which were stained with iodine -thus he considered them related to starch-, and were similar to those previously mentioned inclusions. Although we nowadays know Virchow was actually observing *corpora amylacea*, which are mostly composed of glycosamynoglycans (Sakai et al. 1969; Cohen et al. 2000), the chemical analysis of "amyloid" materials made first by George Budd and later by Carl Friedreich and August Kekulé lead them to conclude these substances were mostly "albuminous", this means of a proteinaceous nature (Sipe & Cohen 2000).

Although the polypeptidic nature of amyloid substances had already been established in the 19th century, and despite of the progress made during the 20th century towards the classification of the different diagnosed amyloidoses, it was long believed that amyloid was rather a concrete substance of possible unspecific degenerative origin (Westermark 2005). It was not until the 1970's that the biochemical characterization of different types of amyloids allowed to establish that each of them was primarily composed of a specific kind of protein or variants of the same protein. Nowadays, the term "amyloid" is employed to denominate a type of essentially proteinaceous extracellular aggregates, which are associated to more than 40 human pathologies (Table 1); although amyloids with a specific biological function have been described in several species, including humans (Maeda et al. 1992)

Table 1.- Amyloid-Forming Proteins Associated to Human Disorders

Protein	Disease	Conformation
α,-synuclein	Parkinson's disease Dementia with Lewy bodies	Intrinsically disordered
β2-microglobulin	Hemodialysis-related amyloidosis	All-β
β2-microglobulin variants	Systemic amyloidosis	All-β
γ-crystallins	Cataract	All β, γ-crystallin like
ABri precursor protein variants	Familial British dementia	Intrinsically disordered
ADan precursor protein variants	Familial Danish dementia	Intrinsically disordered
Amyloid β precursor protein	Alzheimer's disease	Intrinsically disordered
Androgen receptor with polyQ expansion	Spinal and bulbar muscular atrophy	α-helical (nuclear receptor ligand binding domain)
Apolipoprotein AI variants	Systemic amyloidosis	Intrinsically disordered
Apolipoprotein AII variants	Systemic amyloidosis	-

Table 1.- Amyloid-Forming Proteins Associated to Human Disorders (continued)

Protein	Disease	Conformation
Apolipoprotein AIV	Systemic amyloidosis	-
Ataxins with polyQ expansion	Spinocerebellar ataxias, several types	
TATA-box binding protein with polyQ expansion	Spinocerebellar ataxia type 17	α+β (residues 159-339)
Atrial natriuretic factor	Atrial amyloidosis	Intrinsically disordered
Atrophin-1 with polyQ expansion	Dentatorubral-pallidoluysian atrophy	-
Calcitonin	Medullary carcinoma of the thyroid	Intrinsically disordered
Corneodesmin	Localized amyloidosis	-
Cystatin C variants	Icelandic hereditary cerebral amyloid angiopathy	α+β
Fibrinogen α-chain variants	Systemic amyloidosis	-
Galectin 7	Localized amyloidosis	-
Gelsolin variants fragments	Finnish hereditary amyloidosis	Intrinsically disordered
Huntingtin with polyQ expansions	Huntington's disease	Intrinsically disordered
Immunoglobulin heavy chain	Systemic and localized amyloidosis	all-β
Immunoglobulin light chain	Systemic and localized amyloidosis	all-β
Insulin	Injection-localized amyloidosis	all-α
Islet amyloid polypeptide	Type II diabetes	Intrinsically disordered
Keratin	Cutaneous lichen amyloidosis	-
Kerato-epithelin	Hereditary lattice corneal dystrophy	-
Lactadherin fragment	Aortic medial amyloidosis	-
Lactoferrin	Corneal amyloidosis associated with trichiasis	α+β
Leukocyte chemotactic factor-2	Systemic amyloidosis	-
Lung surfactant protein C	Pulmonary alveolar proteinosis	-
Lysoyme variants	Systemic amyloidosis	α+β
Odontogenic ameloblast-associated protein	Calcifying epithelial odontogenic tumors	-
Prion protein	Spongiform encephalopaties	Intrinsically disordered (residues 1-120) α-helical (residues 121-230)
Prolactin	Pituitary prolactinomas	All-α
Semenogelin 1	Localized amyloidosis	-
Serum amyloid A protein fragments	Sistemic amyloidosis, Familial Mediterranian fever	All-α
Superoxide dismutase 1	Amyotrophic lateral sclerosis	All-β
Tau protein	Frontotemporal dementia with Parkinsonism	Intrinsically disordered
Transthyretin	Senile systemic amyloidosis	All-β
Transthyretin variants	Familial amyloidotic polyneuropatry	All-β

adapted from (Uversky & Fink 2004; Chiti & Dobson 2006; Invernizzi et al. 2012; Sipe et al. 2012)

1.4.b.- Determinants of Protein Aggregation

Amyloid aggregates are characterized by a particular fibrilar morphology under **transmission electron microscopy (TEM),** which is highly ordered, compact, stable and unbranched. This type of structures are denominated amyloid fibrils, their diameters range within tens of nanometers and their longitude can reach several micrometers; mature fibrils can further associate laterally to form fibers. Amyloid fibrils have been shown to share a common molecular architecture composed of the **cross-β sheet supersecondary structure**, where parallel β-sheets extend with their strands facing to each other and perpendicular to the fibril axis; such a conformation possesses a characteristic X-ray diffraction pattern (Sunde & Blake 1997). This particular structure provides amyloid fibrils with the capability of binding certain chemical compounds like **Thioflavin-T (Th-T)** and **Congo Red (CR)**, whose spectral properties change upon binding to amyloid fibrils, therefore serving as probes to test the amyloid-like nature of protein aggregates (Nilsson 2004; Hawe et al. 2008). Aside from TEM, fibril morphology can also be detected employing **atomic force microscopy (AFM)** and the conformational change from the native conformation may be followed with **Fourier-transform infrared spectroscopy (FTIR)** by monitoring the appearance of the characteristic intermolecular β-sheet band at ≈1625 cm^{-1}, or with **circular dichroism (CD)** by following changes of the typical negative β signal at ≈217 nm (Figure 5).

The polypeptides involved in the formation of abnormal amyloid-like protein deposits are neither related in sequence, nor in native conformation – some of them being predominantly unstructured (i.e. the amyloid β peptide and α-synuclein) while others are compact globular proteins in their native state (i.e. transthyretin, superoxide dismutase 1 and β2-microglobulin) (Chiti & Dobson 2006). On the other hand, aggregation into amyloid-like fibrils similar to those found in protein deposits is not restricted to a discrete number of polypeptides related to certain human pathologies, but has been shown or induced for a large number of proteins, from different organisms, many of them lacking any known association to disease (Guijarro, Sunde, et al. 1998; Stefani & Dobson 2003; Uversky & Fink 2004). Moreover, cross-β structure has also been reported in macroscopically non-fibrilar, apparently amorphous, aggregates (Wang et al. 2010). These findings have led to the consideration that the ability to adopt the cross-β supersecondary conformation, would constitute an intrinsic property of virtually any polypeptide (Fändrich et al. 2001) since backbone-mediated interactions are the strongest contributors towards the acquisition of such conformations (Knowles et al. 2014).

Nonetheless, the experimental analysis of amyloidogenic proteins and peptides has revealed that changes on its amino acid sequence lead to dramatic changes on its tendency to aggregate (Wurth et al. 2002; Dobson 2003b). These results show how the primary structure of proteins plays an extremely relevant role in determining the tendency of polypeptides to form insoluble deposits; this would arise from the impact over such propensity of inherent physical-chemical properties of amino acids such as hydrophobicity, the structural suitability to adopt β-conformation or its mean charge (Chiti et al. 2003). Specifically, the study of polypeptides able to form amyloid-like fibrils but lacking any defined three-dimensional structure on its physiological context, such as the intensively explored Aβ peptides related to the Alzheimer's disease, has allowed to decouple the determinants of protein aggregation from those of the folding of globular proteins, whose driving forces substantially overlap (Jahn & Radford 2008).

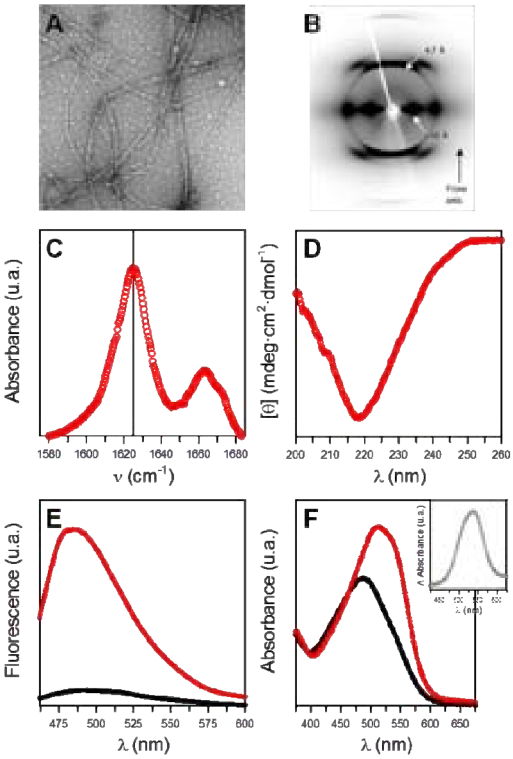


Figure 5.- Classical Experimental Techniques for the Charactherization of Amyloid-like Aggregates. Characteristic properties of amyloid-like aggregates A) fibrillar appearance under TEM, B) characteristic X-ray diffraction pattern indicative of cross-β supersecondary structure (reproduced from (Makin & Serpell 2005) with permission), C) intermolecular b-sheet IR band at ≈1625 cm^{-1}, D) increase in the negative CD signal at ≈217 nm, and change in the spectral properties of amyloid dyes: E) Th-T fluorescence increase and F) CR absorbance shift (the inset indicates the charactheristic band of the difference spectra at ≈540 nm.

1.4.b.I.- Intrinsic Determinants

Hydrophobicity appears to be one of the major determinants promoting aggregation. It has been shown for many proteins or peptides, that mutations substituting polar by non polar residues tend to increase the aggregation propensity and or deposition rate of polypeptides (Esler et al. 1996; Hilbich et al. 1992; Chiti et al. 2003). However, it has been shown that hydrophobicity does not suffice by itself to rationalize the outcome of amino acid replacements on the aggregation potential. Accordingly, attempts to predict protein aggregation propensity solely on the basis of side-chain hydrophobicity have failed (Wurth et al. 2002).

Another relevant intrinsic property of polypeptides is its propensity to adopt specific secondary structure motifs. Consistent with the observation that most protein aggregates share the cross-β supersecondary structure, it has been observed that aggregation is favoured by aminoacids with a higher propensity to adopt β-sheet conformation (Chiti, Taddei, et al. 2002). Moreover, pre-existing β-strands in protein structures also contribute to the tendency of polypeptides to aggregate (Castillo & Ventura 2009). Consequently, aminoacids disfavouring β-conformation, such as Pro and Gly, have been reported to largely disrupt aggregation-prone sequence stretches (Steward et al. 2002; Wood et al. 1995; Parrini et al. 2005).

As a physical-chemical property which can be globally regarded as opposed to hydrophobicity, amino acid charge is known to prevent or disrupt protein deposition (Wurth et al. 2002). As well, the net charge of the polypeptide also influences aggregation (Chiti, Calamai, et al. 2002; Chiti et al. 2003); the higher a protein net charge, the greater the repulsion between individual protein molecules and the lower their chances to establish intermolecular contacts.

The aforementioned factors may be regarded as intrinsic determinants arising from individual properties of amino acids. Nonetheless, the linear combination of their properties along the primary structure has a cooperative impact over the aggregation propensity. It has been observed that the consecutive occurrence of three or more hydrophobic residues is clearly disfavored in nature (Schwartz et al. 2001). In a similar way, the combinatorial design of amyloidogenic proteins has shown how polypeptidic patterns alternating apolar and polar aminoacids favor amyloid formation (West et al. 1999). Remarkably, it has been shown that this patterns are less frequent in natural proteins than it would be expected by random chance (Broome & Hecht 2000).

The failure of protein aggregates to adopt regular macroscopic assemblies hampered, for many years, the possibility of obtaining a detailed description of the structure of amyloids at an atomic level. Fortunately, the development of new techniques, such as solid state NMR (ss NMR) (Petkova et al. 2002) or microcrystallitazion of amyloidogenic peptides (Makin et al. 2005; Nelson et al. 2005) has allowed to unveil the molecular detail of amyloid formation for certain proteins and short peptides. The solved structures provide an outstanding framework to rationalize the intrinsic determinants of protein aggregation. Many of the solved structures correspond to an extended β-sheet whose β-strands are parallel to the axis of the fibril. In these β-sheets, hydrophobic residues are protected from the solvent by establishing interactions with other residues of β-strands in the opposite β-sheets, while polar residues are exposed to the solvent. The geometry of the β-conformation allows the side chains of contiguous residues to point in opposite senses, so this explains how alternation of non polar and polar residues in the primary structure facilitates amyloid formation.

The sequence of a polypeptide determines its folding to a defined three-dimensional structure, which, under physiological conditions, would correspond to its native structure. Therefore,, mutations in the primary structure may affect both the structure and the stability of a protein (Chiti, Calamai, et al. 2002). These variables affect the propensity and the rate to which a polypeptide may aggregate; therefore native conformation and its stability may be considered intrinsic determinants of protein aggregation. Furthermore, these parameters are also intimately related to the mechanisms by which polypeptides aggregate, which will be considered in more detail in section 1.2.c.

1.4.b.II.- Environmental Determinants of Protein Aggregation

The extrinsic determinants of protein aggregations refer to a set of variables defining the environment of the polypeptide chain, which can affect the tendency of a protein to form deposits, since they are able to modulate the intrinsic factors governing aggregation. The most relevant extrinsic determinants are the pH and the ionic strength of the solution, together with the temperature of the system (DuBay et al. 2004). These variables may affect both the kinectics and the thermodynamics of the aggregation into amyloid-like structures, and can, subsequently, influence the assembly and the macroscopic structure of the aggregated species, thus being important determinants of the polymorphism the aggregates of a given protein sequence may present.

The pH directly influences the protonation state of amino acid sidechains and, accordingly, modulates physical-chemical properties such as its polarity and net charge, which, as described above, posses a strong relevance in the aggregation propensity of polypeptides. In the same way, pH also influences the net charge of the protein, thus modulating electrostatic repulsion between individual molecules and, subsequently, the probability of the stability of the intermolecular interactions required for the formation of amyloid-like structures. Ionic strength has a similar role in modulating aggregation since its increase allows to shield amino acid sidechain charges and, therefore, to decrease the repulsion between polypeptide molecules (Morel et al. 2010).

As an intrinsic determinant, the conformational stability of a polypeptide may also be altered by external factors. For those proteins which adopt a defined three-dimensional structure in its physiological context, variables such as temperature or pH my alter the network of interactions stabilizing their native conformations, thus allowing polypeptides to populate partially or globally unfolded states from where aggregation might take place more easily.

1.4.b.III.- Specific Regions Determining Protein Aggregation

The intrinsic determinants described before inform us about specific properties of the amino acid sidechain either favoring or disfavoring protein aggregation. The linear combination of such properties within the primary structure plays a major role in protein aggregation. However, it has been observed that not all the polypeptide sequence has the same importance in defining its propensity to aggregate. There exist small amino acid stretches within protein sequences which promote and guide protein aggregation into amyloid like structures (Ventura et al. 2004; Ivanova et al. 2004). These short fragments, generally referred to as **aggregation-prone regions** (APRs) or **"hot-spots"**, are characterized by an enrichment in hydrophobic, both aliphatic (Val, Leu, Ile) and aromatic (Phe, Trp, Tyr), residues (Rousseau et al. 2006). The analysis of the structural models for some amyloids (Ritter et al. 2005; Krishnan & Lindquist 2005; Kajava et al. 2005) also allow to rationalize the reason why APRs direct the formation of amyloid-like structures, since the cross-β arrangement in the core of amyloid fibrils does only strictly require the minimum participation of a single β-strand per molecule, and the rest of the polypeptide may remain exposed to the solvent even when "attached" along the fibril. APRs are usually located within or substantially take part of the hydrophobic cores of the native state of proteins (Linding et al. 2004) and also frequently map to protein-protein interaction surfaces of protein adopting stable quaternary structure (Pechmann et al. 2009; Castillo & Ventura 2009), which prevents APRs from establishing aberrant intermolecular contacts.

The examination of APRs in the context of entire proteins has also revealed how these stretches are usually flanked by charged residues (Asp, Glu , Lys , Arg), whose function would be to hamper intermolecular interactions between APRs in the event they become exposed by providing repulsive charge or by residues acting as β-sheet breakers, like Pro (Rousseau et al. 2006; Reumers et al. 2009).

1.4.c.- Mechanisms of Protein Aggregation

While the appearance of protein aggregates may differ at the macroscopic level, evidence from an array of different experimental techniques indicates that the final stages of aggregation are enriched in β structure (Wang et al. 2010). In particular, X-ray diffraction analysis of either amyloids, amyloid-like fibrils, inclusion bodies or amorphous aggregates present a characteristic pattern (Figure 6) representative of cross-β supersecondary structure (Sunde et al. 1997; Wang et al. 2008; Ramshini et al. 2011). Therefore, ultimately, a mechanism for protein aggregation should describe the conversion to this, apparently universal, cross-β superfold from the physiological native state of a polypeptide.

The analysis of the aggregation kinetics of different proteins early showed to follow a sigmoidal behavior employing a variety of techniques (Morris et al. 2009). Although different models had been proposed to explain the mechanism of amyloid formation based on the analysis of the prion protein such as the **Templated Assembly** (Griffith 1967) or the **Monomer-directed Conversion** (Prusiner 1982) models, the one which has proven more successful in describing the common triphasic nature of formation of amyloid-like structures is the **Nucleation-dependent Polymerization** model (Jarrett & Lansbury 1993). According to the latter, three main stages may be differentiated during the assembly reaction. The first corresponds to the nucleation of aggregation, during which protein monomers associate to form oligomers that will then act as nucleus to propagate the aggregation reaction; this constitutes a thermodynamically unfavorable process and is characterized by a lag phase characterized by a

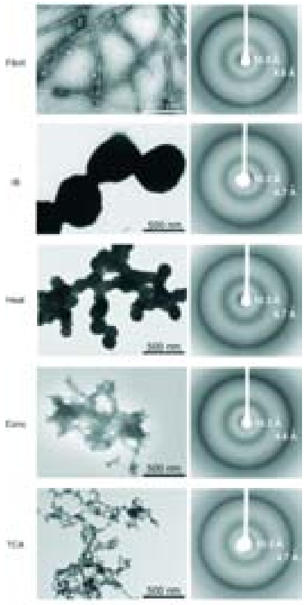


Figure 6.- X-ray Diffraction Pattern of Different Types of Protein Aggregates. The X-ray diffraction patterns of different types of aggregates formed by the N-terminal domain of the hydrogenase maturation factor HypF (amyloid-like fibrils, inclusion bodies -IB-, and aggregated induced by heat, concentration -Conc- or trichloroacetic acid -TCA-) all present the charactheristic reflections of the cross-β supersecondary conformation at ≈4.7 Å and ≈10.2 Å. Reproduced from (Wang et al. 2010) with permission.

slow increase, if any, in the signal of aggregated species (Figure 7). Since the nucleation phase involves monomer self-association, protein concentration has been shown to deeply affect the nucleation stage (Harper & Lansbury 1997; Cohen et al. 2012) due to the concomitant raise in the probability of interaction between polypeptide units as concentration increases. Once a sufficient amount of nucleus is accumulated, the thermodynamically favorable propagation of amyloid-like structures increases exponentially and ultimately reaches saturation.

A further development of the Nucleation-dependent Polymerization model, the **Nucleated Conformational Conversion** mechanism (Serio 2000), is currently the most widely accepted model to describe the formation of amyloid-like structures. In contrast with its predecessor which stated that polypeptide monomers would experience a conformational transition to an aggregated-like monomeric state before they could associate into oligomers, the latter model considers that aggregating proteins experience changes of its native state that lead them to populate aggregation-prone conformations, these conformations can establish favourable interactions with other aggregation-prone units in order to form early oligomers. Those early oligomers are not expected to possess a specific stoichiometry but are rather characterized by equilibria between different multimeric species (Figure 8); within them, a rate-limiting conformational conversion of the monomeric subunit takes place to adopt cross-β structure, yielding a mature oligomeric form (Orte et al. 2008; Bemporad & Chiti 2012). These late oligomers possess the structural suitability to proceed with a fast elongation phase into protofibrilar and later fibrilar species. The exhaustion of oligomeric intermediates leads to a stationary phase, which is characterized by a *plateau* in the signal of the aggregation kinetics. Fibrils shall not be regarded as inert, static end products during this late stage; on one side they can associate laterally with other fibrils to yield mature fibers composed of several protofilaments (Sunde & Blake 1997; Serpell et al. 2000; Jiménez et al. 2002), as well fibrils have been shown as dynamic structures which undergo a continuous molecular recycling by oligomer dissociation and reassociation within the fibril (Carulla et al. 2005) and can also experience fragmentations (Tanaka et al. 2006).

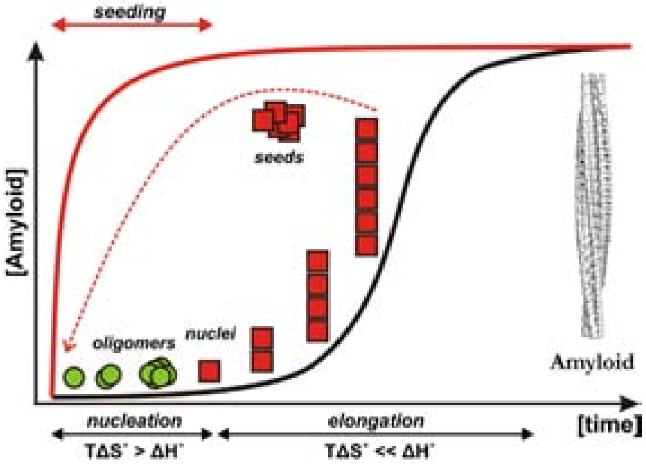


Figure 7.- Typical Kinetic Profile of Aggregation according to a Nucleation-dependent Mechanism. Three stages are differentiated during the course of a nucleated-dependent aggregation reaction: a lag phase corresponding to nucleation of aggregation, a second elongation phase where oligomeric species assemble to grow into amyloid-like fibrils, until finally reaching a plateau where fibrils do not grow further. Seeding the reaction with preformed mature aggregated species significantly reduces the lag phase by abrogating the thermodynamically unfavorable nucleation stage. Reproduced from (Eichner & Radford 2011) with permission.

Consistent with the observation of the nucleation stage being the rate-limiting step of the aggregation into amyloid-like fibrils, it has been shown that the presence of preformed nucleus accelerates fibrillation, reducing the lag phase, or even abrogating when a sufficient amount of preformed nucleus is added (Harper & Lansbury 1997; Cohen et al. 2012). This phenomenon is denominated seeding and the preformed nuclei are therefore referred as seeds, it bares the same rationale behind the seeding procedure with preformed crystals employed to foster protein crystallization in structural studies (Wolde 1997). It has been observed that the seeding efficiency decreases as the sequential identity between the seed and the soluble monomer is lowered (Krebs et al. 2004), this is mainly attributable to the self-complementary required for an effective packing of APRs into the cross-β conformation, and is consistent with the recurrent observation of amyloids being highly homogeneous in composition, even when retrieved from tissue deposits, with a single polypeptide constituting its major component. Moreover, not only the composition but the linear arrangement of amino acids in the primary sequence of APRs is crucial for effective seeding, since either reverted or scrambled versions of a given sequence are unable to cross-seed aggregation into amyloid-like fibrils (Sabaté et al. 2010).

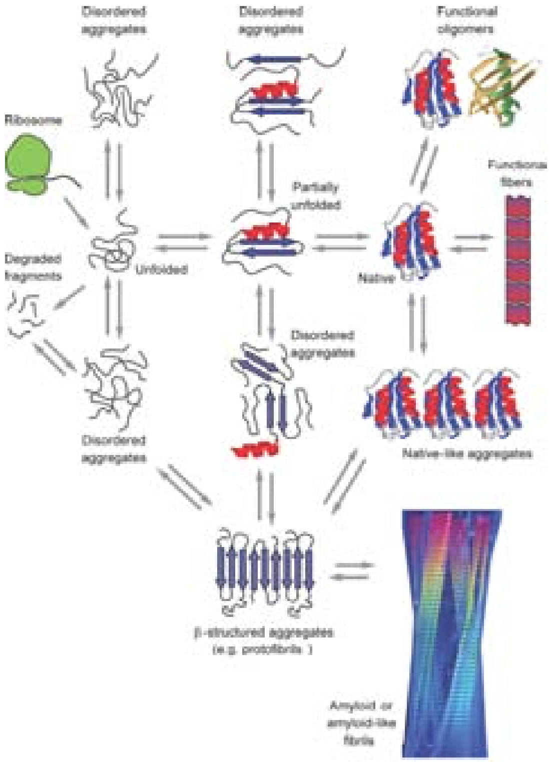


Figure 8.- Schematic Representation of Conformational Diversity within the Species Populated during Aggregation. The aggregation of a polypeptide may be initiated from different conformational states, and during their aggregation they can populate highly diverse ensembles of conformations which readily interconvert among them, depending on their relative thermodynamic and kinetic stabilities. Reproduced from (Chiti & Dobson 2006).

Despite the Nucleated Conformational Conversion mechanism described before allows to model the kinetics of amyloid-like structure formation by most aggregation-prone proteins from a general perspective, the strikingly divergent tertiary and quaternary structures adopted by those proteins able to form amyloids indicates that no general model would account for the specific mechanism of conversion. Several models have been proposed to explain the conformational conversion required for amyloid-like aggregation based on both the structure and the conformational stability of polypeptides (Nelson & Eisenberg 2006).

A first model relies on the observation that many amyloids are composed of proteins or fragments of proteins that lack a defined three-dimensional structure in its physiological environment, like amyloid β or α-synuclein. According to this **Intrinsically Unstructured** model, aggregation-prone polypeptides without a defined three-dimensional conformation can establish intermolecular interactions in their native (unstructured) conformation. However, this does not imply this class of proteins can readily form amyloid-like structures, they also must undergo a transition to a conformation compatible with amyloid structure (Cremades et al. 2012), as this is supported by the fact a lag phase is also observed in the aggregation kinetics of this kind of proteins.

The establishment of intermolecular interactions between APRs within intrinsically unstructured conformations can be rationalized in a simple manner, but the way APRs protected by the native state of globular proteins get to interact to form amyloid-like structures is not straightforward. A different mechanism to explain the conversion of globular proteins into amyloid-like structures, the **Refolding** model, proposes substantial unfolding of the native conformation is required for APRs to become exposed and being able to establish effective interactions leading to conformational conversion into the cross-β fold. Substantial perturbation of the native state required to populate such largely unfolded conformations, as those observed at the starting point of the aggregation into amyloid-like fibrils induced for proteins like the SH3 domain or myoglobin (Guijarro, Sunde, et al. 1998; Fändrich et al. 2003), would arise from the impact of mutations or changes in the environmental conditions in the global stability of the protein.

However, overcoming the usually large unfolding barrier, as required by the Refolding model, does not seem feasible for a vast majority of aggregation-prone globular proteins, simply by means of single substitutions or normal changes in its physiological environment. Indeed, aggregation into amyloid-like fibrils has been reported for proteins departing from native-like conformations like acylphosphatase and lysozyme (Chiti & Dobson 2009), which in some cases are even able to retain a certain functional activity (Plakoutsi et al. 2004; Bemporad et al. 2008). Another mechanism of conformational conversion, the **Gain-of-Interaction** model, has been proposed to rationalize aggregation into amyloid-like structures without requiring crossing of the unfolding barrier. According to this model, local conformational perturbations of the native state would be sufficient to expose previously protected APRs, in a way they can now establish non-functional intermolecular contacts with other polypeptide unit displaying similarly exposed stretches. These locally disordered states can be reached directly through fluctuations of the native conformation, while the remainder of the native structure may well be left unaltered and the protein molecule may even retain its physiological activity to a certain extent. The precise mechanism through which this gain-of-interaction is achieved depends on several factors such as the fold and size of the protein, the location of APRs within the native conformation and the quaternary structure of the polypeptide. Three different subtypes among the Gain-of-Interaction model have been proposed to account for such variables.

The simplest mode of gain-of-interaction implies exposed APRs directly interact with each other to form the **cross-β spine** of the fibril, while the remainder of the polypeptide is left “hanging” attached to this fibril spine. This model requires a certain self-complementarity between the sidechains of the polypeptide APRs, which will constitute the future strands of the extended β-sheet that defines the spine of the fibril. Extension of the fibril results from the successive stacking of APRs to form the β-sheet through the establishment of backbone hydrogen bonds. The analysis of the X-ray structures derived from microcrystals of fibril-forming peptides has allowed to define the geometrical requirements for APRs to build cross-β spines following a complementarity scheme that has been termed as *steric zipper* (Sawaya et al, Nature 2007)

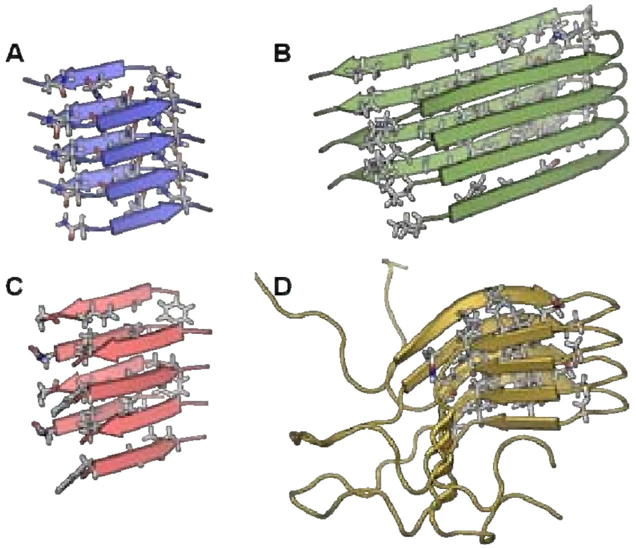


Figure 9.- Steric Zippers Formed by the cross-b Core in Experimentally Resolved Amyloid Structures. Ribbon representation of the cross-β core of the fibrils formed by A) the NNQQNY peptide form the *S. cerevisiae* Sup35 prion forming domain (PDB 2omm), B) the Ab42 peptide (PDB 2beg), C) a fragment of the human islet amyloid polypeptide (PDB 2kib), and D) the prion forming domain of HET-s from *P. anserina* (PDB 2rnm). The sidechains of the residues conforming the steric zipper are shown in light gray.

Another mode through which polypeptides may undergo gain-of-interaction consists in the rearrangement of APRs to the surface of the protein, so a novel homooligomerization surface is generated, therefore allowing **direct stacking** of protein monomers as a mechanism of fibril extension (Elam et al. 2003). This model of conformational conversion has been proposed for the aggregation all-β proteins with a native oligomeric quaternary structure, such as superoxide dismutase 1 (SOD1) (Elam et al. 2003) or transthyretin (TTR) (Serag et al. 2002; Olofsson et al. 2004). These proteins posses a β-sandwich fold and display protein-protein interaction surfaces that allow them to form native, soluble oligomeric structures. In both cases, either partial or global preservation of native dimerization surfaces has been proposed, but the emergence of interaction surfaces in the opposite side of the molecule, able to establish contacts with identical regions of neighboring molecules, allows protein oligomerization to extend linearly into amyloid-like fibrilar structures where the strands of the mostly native β-sandwich are disposed perpendicular to the fibril axis, thus mimicking the cross-β assembly.

Finally, a third alternative mode for proteins to experience gain-of-function is **3D domain swapping** where at least two identical protein entities exchange part of their structural elements to yield stable oligomeric species, as it has been show for cystatin C (Janowski et al. 2001). Extension of the oligomer size may be compatible with growth into an amyloid-like fibril when the swapping mechanism results in the formation of β-conformation able to dock in a cross-β spine fashion, or when it involves the successive structural exchange between adjacent polypeptide units (Sambashivan et al. 2005).

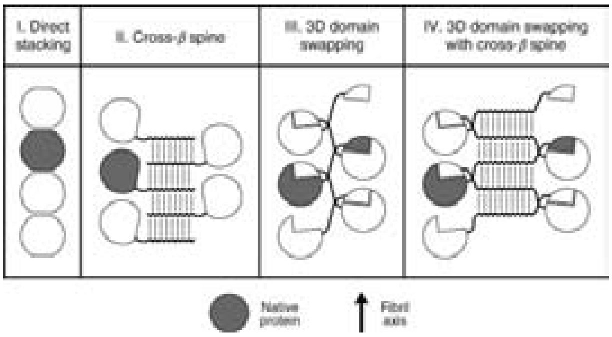


Figure 10.- Schematic Representation of the Different Modes of the Gain-of-Interaction Mechanism for Amyloid Assembly. Reproduced from (Nelson & Eisenberg 2006) with permission.

1.5.- Protein Folding and Aggregation in the Cell

Although the protein deposits related to different human pathologies are know to be mainly composed of amyloid fibrils, these are only observed in vivo for secreted proteins or fragments of proteins whose aggregation takes place in an extracellular environment. Inside the cell, virtually any biochemical process is subject to tight regulation, and the ensurance of proper polypeptide folding, as well as the management of misfolded species are not exceptions to this regulatory pressure.

1.5.a.- Cellular Safeguards for Correct Protein Folding

The nascent polypeptide chain remains protected from spurious interactions as it emerges from the ribosome thanks to its interaction with ribosome-binding chaperones such as the triggering factor (TF) in prokaryotes, the ribosome-associated complex (RAC) in *Saccharomyces cereviase*, the nascent- chain-associated complex (NAC) in archea and eukaryotes and other ribosome-associated chaperones that have been identified in mammals (Preissler & Deuerling 2012). They generically bind and shield hydrophobic patches in the newly synthesized chain from solvent. In bacteria, the ATP dependent release of TF from the nascent chain either allow folding to the native state or its transfer to chaperones acting downstream of the ribosome. TF also catalyzes the cis/trans isomerization of peptidyl-prolyl bonds, which can be a rate-limiting step in protein folding.

Proteins that cannot spontaneously fold into the native conformation after being released from ribosome-binding chaperones are transferred to the canonical Hsp70 chaperone system (Kim et al. 2013). This class of chaperones may directly assist the folding of proteins to the native conformation or, if this is not achieved, its transfer to other chaperone systems by binding to short hydrophobic stretches of extended polypeptides which are flanked by charged residues (Rüdiger et al. 2000; Tartaglia et al. 2010), a binding pattern that deeply resembles the sequential features of APRs, as discussed above. Chaperones from the Hsp40 family cooperate with the Hsp70 class by binding non-native proteins and directly mediating their transfer to Hsp70 (Kampinga & Craig 2010).

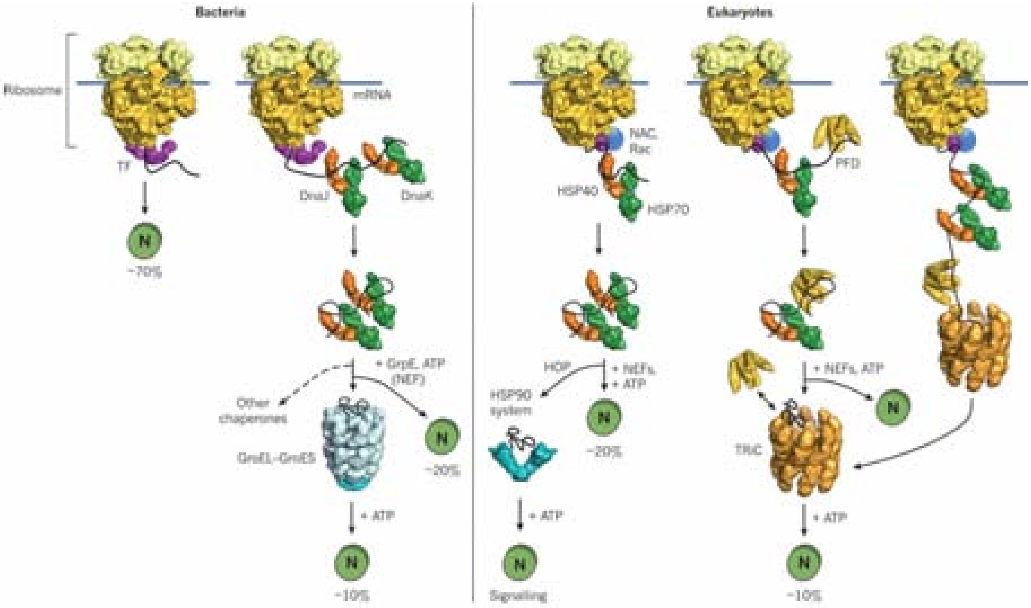


Figure 11.- Overview of the Prokaryotic and Eukaryotic Machineries Assisting Protein Folding in the Cytosol. The nascent chain may binds to ribosome-associated chaperones at exit tunnel of the ribosome. If it requires further assistance for folding, it may be transferred to chaperones the Hsp70 system. When the native conformation of the polypeptide is not achieved neither at this stage, the protein can be subsequently derived to members of the Hsp60/chaperonin or Hsp90 chaperone systems. Reproduced from (Hartl et al. 2011) with permission.

When the action of Hsp70 chaperones does not suffice to help proteins fold to its native structure, these are further transferred to the Hsp60 or the Hsp90 chaperone systems (Kim et al. 2013). The Hsp60 class, also known as chaperonins, form large complexes with double a ring structure; creating a cavity where isolated proteins molecules can fold without the risk of establishing aberrant interactions (Horwich et al. 2007). During this process the cavity experiments a conformational shift, which changes its inner character from non-polar to polar triggering the dissociation of the non-native polypeptide from the wall of the cavity. The released protein would then be free to reinitiate folding as an isolated molecule in a wáter-like environment. It has been estimated that about 10-15%,and 5-10% (Yam et al. 2008; Kim et al. 2013), of nascent polypeptides in bacteria and eukaryotes, respectively, require the assistance of the Hsp60 chaperones to fold to their native conformation.

The Hsp90 chaperone system is involved, on the other hand, in the maturation of eukaryote signaling proteins (Taipale et al. 2010). Protein substrates are transferred from the Hsp70 class to the Hsp90 system that entraps polypeptides in a molecular clamp conformational transition which promotes the proper folding of its client polypeptides (Southworth & Agard 2011; Li et al. 2012).

1.5.b.- Cellular Management of Misfolded Species

When the chaperone network fails to assist the correct folding of a polypeptide, misfolded species may be generated. It has been proposed that certain E3 ubiquitin ligases might be able to recognize misfolded proteins and tag them with ubiquitin polymers linked through Lys63, which serves as a signal for degradation by the ubiquitin-proteasome system (UPS) (Goldberg 2003).

Under certain circumstances the machinery in charge of assisting the proper folding of proteins or its degradation may be overrun; i.e. due to mutations exacerbating the aggregation propensity of polypeptides, errors during normal protein biogenesis or because of ageing under environmental stress (Tyedmers et al. 2010). In such a case, misfolded species may start to accumulate and form protein aggregates. However, the formation of polypeptide aggregates does not occur at random but is also subject to regulation in eukaryotic cells. In yeast, misfolded species are sequestered into two specific kinds of aggregated deposits (Kaganovich et al. 2008). On one side, some proteins are stored in a structure adjacent to the nuclear membrane, which is termed the Juxtanuclear Quality-control Compartment (JUNQ), this is regarded as temporary deposition site for proteins that can be further refolded or degraded. In fact, many proteins in JUNQ possess the poly ubiquitin degradation tag. A different storage structure is the Insoluble Protein Deposit (IPOD), located adjacent to the vacuole, and where terminally misfolded proteins that cannot be further refolded or degraded by the UPS accumulate. These deposits are removed from the cell by autophagy.

Structures similar to JUNQ and IPOD have been reported in mammalian cells (Kaganovich et al. 2008), nonetheless, a specialized sequestration compartment exist in mammals, the aggresome, which is located in the perinuclear space, at the microtubule organizing centre (MTOC) (Johnston et al. 1998; Robinson 1976), where misfolded species converge trough an active microtubule-dependent transport.

Protein rescue from aggregate deposits is known to involve the action of a bichaperone system (Glover & Lindquist 1998; Weibezahn et al. 2004). First a complex of Hsp70 and Hsp40 class chaperones extracts a polypeptide chain of the aggregate (De Los Rios et al. 2006), the action of the Hsp70 chaperon may suffice to properly refold the protein (Lewandowska et al. 2007)but when it does not, the polypeptide is transferred to a different type of chaperone belonging to the Hsp100 system (Zietkiewicz et al. 2004; Haslberger et al. 2007). This class of chaperones acts by threading the substrate chain through a central pore in its structure, this mechanical phenomenon is led by aromatic residues within the pore. Threading proceeds until the Hsp100 encounters a properly folded region of the protein end then releases the substrate to allow its non-assisted refolding.

The clearance of aggregated deposits may also proceed through the action of the AAA+ proteases directly on aggregate molecules (Tyedmers et al. 2010). AAA+ proteases activity is blocked by the presence of the Hsp70-Hsp40 complex, which restricts the protease accessibility to the substrate, thus suggesting that protein refolding would be preferred before irreversible degradation occurs. However, it has been observed that in higher eukaryotes, the elimination of aggregated intracellular deposits seems to proceed preferentially through selective autophagy mechanism (Tyedmers et al. 2010). Following this mechanism, aggregates are enveloped in an autophagosome that later fuses to a lysosome, where hydrolases will degrade misfolded polypeptides (Rubinsztein 2006; Kirkin et al. 2009).

1.6.- Pathogenic Consequences of Protein Aggregation

The observation that amyloid fibrils are the major constituents of the protein deposits associated to several human disorders such as the Alzheimer´s disease led to the initial consideration that this protein conformation might be the causative agent of disease in conformational disorders. However, the research efforts devoted during the past 15 years towards the characterization of the aggregation process of several proteins involved in such kind of diseases has changed the view regarding the mechanism of pathogenicity.

Although the analysis of the toxic activity that may be gained by species populated along the aggregation pathway has attracted much attention, a straightforward detrimental effect of the conformational misfolding experienced by many polypeptides its the decay of its cellular function and/or location. Pathogenic consequences of this **loss-of-function** may arise directly from the collapse of the protein activity or as a result of the accumulation of substrates or deprivation of products or related metabolites either upstream or downstream from protein function (Gregersen 2006).

Even when the loss of the native conformation does not suffice *per se* to completely abrogate protein activity, the induction of PQC machinery sequestration pathways may suffice to promote a loss of function at the cellular places where this activity is required. In any case, a toxic gain-of-function, acting either independently or in concert with loss of protein activity, cannot be neglected as a pathogenic mechanism derived from protein aggregation (Winklhofer et al. 2008).

The accumulation of large protein deposits, as observed in several systemic amyloidosis can lead to mechanical impairment and disrupt tissue architecture (Merlini & Bellotti 2003; Eisenberg & Jucker 2012). However, it is widely accepted nowadays that, among the species populated during the process of aggregation, amyloid fibrils are not the most toxic for the cell. In fact, a large body of evidence indicates that small oligomeric intermediates formed along amyloid aggregation are the most toxic species (Bucciantini et al. 2002; Kayed et al. 2003). The elevated polymorphism and multiple equilibria established between the myriad of oligomeric intermediates populated during protein aggregation (Kodali & Wetzel 2007; Stefani 2010) has impaired the detailed characterization of the toxicity determinants in these metastable species. Two parameters have been extensively reported to correlate with oligomer cytotoxicity, though. On one side cellular toxicity has been found to decrease as the size of oligomer species decreases (Bemporad & Chiti 2012). On the other hand, increased exposure of hydrophobic patches at the protein surface, as assessed by 1-anilinonaph-thalene 8-sulfonate (ANS) binding, also correlates with higher toxicity (Bolognesi et al. 2010; Mannini et al. 2014). These observations can be rationalized in the context of the nucleated conformational conversion mechanism of amyloid assembly: initially misfolded polypeptides may exhibit hydrophobic stretches previously protected within the native conformation, as monomeric or low molecular mass oligomeric species coalesce into intermediates with higher mass through the establishment of unspecific interactions, hydrophobic exposure is reduced by means of shielding resulting from intramolecular contacts of hydrophobic nature and the decrease of the surface to volume ratio as the oligomer size increases. Additionally, increasing oligomer size would be paired to the reduction of its diffusivity within the cell, thus constraining the chances of establishing spurious interactions with other cellular components.

The precise molecular mechanisms accounting for oligomer citotoxicity are still poorly understood. Several mechanisms have been proposed, ranging from deleterious interactions altering the function of other cellular proteins (Cissé et al. 2011) or disrupting membrane integrity (Kremer et al. 2000; Campioni et al. 2010; Reynolds et al. 2011) to the formation of pores within cellular membranes by annular oligomers (Last et al. 2011) which would disrupt the membrane potential and cellular homeostasis.

Consistent with these findings, the view has emerged that the formation of amyloid fibrils as well as the sequestration of misfolded species by the PQC machinery play a protective role in front of the exacerbated toxicity presented by oligomers. Even though, amyloid fibrils still retain a certain cytotoxic potential (Bemporad & Chiti 2012), which has been attributed to its dynamic nature, exemplified by the molecular exchange between fibrils and oligomeric species (Carulla et al. 2005). In this line, diminished stability of the amyloid fibril end product would result in an increased cellular toxicity (Graña-Montes et al. 2012), likely caused by a shift of the molecular exchange equilibrium towards the accumulation of an increased population of oligomeric intermediates or because of fibril fragmentation leading to secondary nucleation reactions by which monomeric polypeptides convert in novel oligomeric assemblies (Cohen et al. 2013).

An additional mechanism of pathogenesis may be exerted trough a concurrent sequestration of functional proteins, together with misfolded species, by the cellular machinery in charge of the management of aberrantly folded polypeptides (Olzscha et al. 2011). Among the proteins proposed to be subject of this co-sequestration phenomenon, chaperones and proteases are candidates to experience this side-effect, since they directly interact with polypeptides adopting aberrant conformations, which are the primary substrates of sequestration. This secondary sequestration of relevant components of the PQC machinery might result in a reduction of its effective concentration, subsequently leading to a decreased ability of the cell to assist the folding of other proteins different from those initially triggering the PQC system overload (Gidalevitz et al. 2006) and thus in amplification of the population of misfolded species.

Aside from the components of the PQC, its has been proposed that other proteins with relevant functions, such as transcription factors, may also be sequestered within protein deposits leading to cellular malfunction (Olzscha et al. 2011).

CHAPTER 2.- OBJECTIVES OF THE PRESENT THESIS

The works that compose the present thesis are all joined by a common purpose, namely understanding the mechanisms Nature has developed in order to confront and control the risk of protein aggregation. In the last years, the power of the computational tools developed to predict protein aggregation propensity has been employed to analyze large sets of protein ensembles form different organisms, the results turn out to indicate the overall tendency to aggregate in proteomes decreases as the complexity of organisms and its lifespan grows, thus providing evidence for the notion that avoidance of protein aggregation acts as a strong constraint that carves the evolution of protein sequences. At the same time, these analyses have also contributed to identify the determinants of protein aggregation, which, in turn, reveal the variety of strategies proteins have evolved to avoid aggregation.

Despite protein sequences are globally under a strong evolutive pressure to avoid aggregation, the almost ubiquitous presence of APRs within globular proteins argues for the existence of specific functional constraints which exert limits on the selective purification of aggregation-prone stretches from proteins sequences.

The general aim of the present work is to delve deeper into the characterization of the mechanisms evolved by proteins to prevent aggregation, and into the identification of the functional restraints that limit the selective pressure against it. In order to achieve this aim, we have employed different protein model to analyze their aggregative properties. The particular objectives of the specific analysis on each of those models can be summarized in the following points:

- To unravel the role of disulfide cross-linking in the formation and properties of amyloid-like structures.
- To characterize the putative anti-aggregational role of the Nter unstructured extensions present in the SUMO domains.
- To exploit the experimentally determined data available for the *E. coli* model system in order to assess how does the cell regulate the abundance of proteins in relation with their real tendency to aggregate.
- To characterize the tendency to aggregate of kinase proteins, taken as an archetype of a homogeneous and functionally relevant ensemble of proteins expected to be under a significant, yet similar, selective pressure to confront aggregation, in order to identify functional constraints exerting limitations on the evolutionary selection against aggregation.
- To analyze the depositional properties of disulfide-rich domains (DRDs) in order to understand how they confront the risk aggregation when they populate their fully or partially reduced states.
- To extract insights from the former analysis for the understanding of the divergent oxidative folding pathways DRDs may follow.

CHAPTER 3.- MATHERIALS AND METHODS

This chapter is not intended as a broad compendium of the specific methods employed in the published works which compose this thesis, a detailed description can be found in the particular “Matherials and Methods” section of each of the articles. The common experimental techniques employed for the characterization of protein aggregation into amyloid-like structures, which were extensively used throughout these works, have already been introduced in section 1.2.b. Here, a general description is provided of the protein models, specific ensembles of proteins, and datasets of protein properties employed for the analysis of the evolutionary strategies to overcome protein aggregation, which are deeply discussed in the following chapters. As well, since computational tools to predict protein aggregation are the major workhorse in the articles composing this thesis, an overview is presented of the computational methods currently available.

3.1.- Proteic Models

In the following section, a general description will be provided of the different individual proteins, structural or functional ensembles of proteins and protein databases that have been employed to test the anti-aggregational strategies discussed throughout this thesis.

3.1.a.- Protein Domains

3.1.a.I.- PI3-SH3 Domain

Src homology 3 (SH3) domains mediate protein-protein interactions (Pawson 1995; Dalgarno et al. 1997) and have been identified in hundreds of multidomain proteins (Larson & Davidson 2000). Because of their monomeric state, small size and absence of cofactors and catalytic activity, they have been extensively employed as model proteins to study both folding (Grantcharova et al. 1998; Martínez & Serrano 1999; Grantcharova et al. 2000; Ventura et al. 2002) and aggregation (Guijarro, Sunde, et al. 1998; Ventura et al. 2002; Ventura et al. 2004). The SH3 domains share an all-β fold which averages 60 amino acids and is composed of 5 β-strands arranged into two sheets, which form a β-sandwhich with the sheets orthogonal one respect to the other (Figure 12.A). The β-strands in the SH3 domain are linked by characteristic loops which may present variation between the different SH3 domains. β-strands A and B are connected by the RT loop which adopts an irregular antiparallel β-hairpin structure, strands B and C are linked by the n-Src loop charactherized by and 3_{10}-helical conformation at its N terminus, and the strands C and D are connected by a shorter loop denominated Distal loop. In this work we have employed the SH3 domain of the p85α subunit of bovine phosphatidyl-inositol-3-kinase (PI3) which consists of 83 residues and is characterized by an extension of its n-Src loop relative to other SH3 domains (Liang et al. 1996). PI3-SH3 was one of the first globular proteins not associated with any know disease observed to form amyloid-like fibrils, by departing from a particular state populated under acidic conditions (Guijarro, Sunde, et al. 1998). The process of aggregation of this domain is one of the better characterized and it was one of the models employed to establish the "Hot-spot" hypothesis (Ventura et al. 2004), when it was shown that introducing a short stretch of the PI3-SH3 sequence into the α-spectrin SH3 domain, whose wt form does not form aggregates under acidic conditions, could induce its aggregation into amyloid-like fibrils.

3.1.a.II.- SUMO Domains

Small ubiquitin-related modifiers (SUMO) are small monomeric proteins present in all eukaryotic organisms (Geiss-Friedlander & Melchior 2007). They are involved in the regulation of a multitude of important cellular processes (Johnson 2004; Hay 2005; Geiss-Friedlander & Melchior 2007; Wilkinson & Henley 2010; Flotho & Melchior 2013) and exert their function by being covalently conjugated to their protein substrates. SUMO domains all belong to the ubiquitin-like (Ubl) protein superfamily, whose members fold into a mixed five-stranded β-sheet with an extended helix packed against it, and a small α-helix connecting the short β4 and β5, according to a ββαββαβ topology. The ββαββ core of the SUMO structure corresponds to ubiquitin-like or β-grasp superfold, one of the nine most frequent folds (Orengo et al. 1994) which is populated by proteins sharing very low sequence homology. Accordingly, SUMO proteins share a deep structural similarity with Ubiquitin despite their sequential homology being only about 20% at best. More remarkably, they also share with Ubiquitin a similar mechanism of conjugation to their protein substrates which involves the formation of a isopetidic bond between a Cter Gly residue and an acceptor Lys of the substrate, and is catalyzed by the sequential action of a similar system of enzymes (Hay 2005; Gareau & Lima 2010). The main differences between SUMO domains and Ubiquitin reside on the charge distribution at the surface of the proteins, the nature of their ability to form polymeric chains and the presence of a highly flexible unstructured Nter extension in all SUMO domains (Figure 12.B). While Ubiquitin can form polymeric chains through the presence of a specific Lys, this residue is not conserved in the globular Ubl domain of SUMO proteins (Flotho & Melchior 2013). Nonetheless, a consensus SUMOylation pattern is present in the Nter tail of some SUMO domains, which allows for the formation of polySUMO chains (Johnson 2004). SUMO domains are essential for most eukaryotes, although many organisms have evolved different forms (Flotho & Melchior 2013). Only one SUMO form has been found in unicellular eukaryotes and in *C. elegans* and *D. melanogaster*, while several forms are found in mammals and plants. In humans up to four forms have been identified, however SUMO2 and SUMO3 are 95% identical and the role of SUMO4 remains largely unknown since it appears to be incompetent for SUMOylation (Owerbach et al. 2005). Therefore SUMO1 and SUMO2 are considered the predominant forms in human cells.

SUMO domains have been long considered a paradigm of protein solubility (Zuo et al. 2005; Marblestone et al. 2006). Conversely, we have recently been able to show that the destabilization of their native conformation leads to deposition into amyloid-like fibrils (Sabate et al. 2012)

3.1.b.- Protein Ensembles

3.1.b.I.- Ensemble of Protein Kinases (Kinome)

Kinases are one of the most important protein effectors since they are involved in the regulation of virtually any cellular process in all domains of life (Leonard et al. 1998; Manning et al. 2002). Kinases that phosphorylate their protein substrates by catalyzing the transfer of the γ-phosphate of a purine nucleotide triphosphate (ATP or GTP) to a free hydroxyl group of a Tyr, Ser or Thr residue of their polypeptidic substrate, yielding a phosphate monoester, constitute one of the largest families whithin eukaryotic genomes, accounting for about 1.5-2.5% of all their genes (Manning et al. 2002). Therefore, this kind of kinases are usually referred to as eukaryotic protein kinases, alhtough they have been also found in some prokaryotes (Leonard et al. 1998), and their catalytic domains have been shown to share a similar fold, which defines the protein kinase-like superfamily. This fold is composed of two $\alpha+\beta$ domains, namely a smaller Nter domain with a predominantly antiparallel β structure which is responsible for the binding and orientation of the purine nucleotide, and a larger mostly α-helical Cter domain where binding of the substrate and phosphotransfer initiation take place (Hanks & Hunter 1995; Taylor & Kornev 2011) (Figure 12.C). Within this catalytic domain, 12 highly conserved residues have been identified that are considered to play and essential role in kinase function. There are different types of kinases that do not phosphorylate proteins but other kinds of biomolecules, such as lipids, which also share the protein kinase-like fold. The ensemble of kinase proteins that share this specific fold have been termed “typical” kinases (Scheeff & Bourne 2005), in contrast to “atypical” kinases that do not share sequential homology to members of the eukaryotic protein kinases and present substantial structural differences relative to the protein kinase-like fold, largely unrelated kinases are enclosed within this latter group such as His-Asp phosphotransferases, which are common in bacteria (Hanks & Hunter 1995), and kinases that phosphorylate non-proteinaceous substrates.

Aside from those members categorized in the atypical group, protein kinases are classified, based primarily on the sequential similarity of their catalytic kinase domains, in 9 major groups which are further subdivided into 119 families (Manning et al. 2002). Members of all these 9 groups are present in metazoans and 7 of them are also found in yeast. This suggests a remarkable conservation of the functions exerted by kinases, thus making this ensemble of proteins an outstanding model for the evaluation of protein evolution through cross-species analysis.

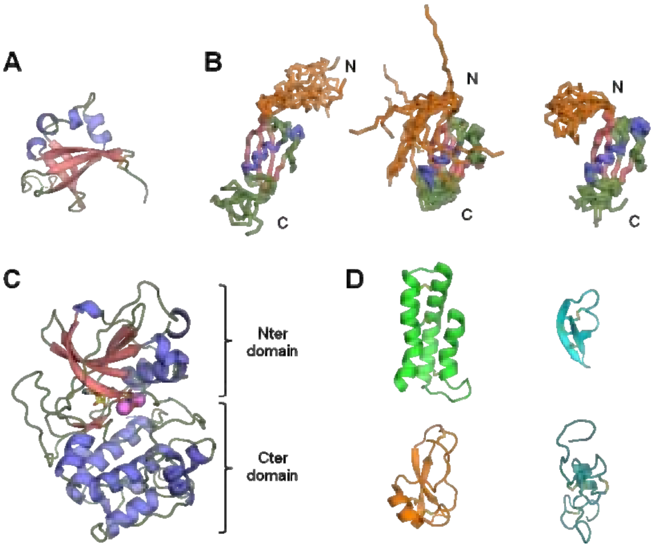


Figure 12.- Structural Representation of the Protein Models Employed in this Work. Cartoon representations of A) the crystal structure of the PI3-SH3 domain (PDB 1pht), B) solution structure conformes of human SUMO1 (PDB 1a5r) and SUMO2 (PDB 2awt) and *Drosophila melanogaster* SUMO (PDB 2k1f). N and C indicate the Nter and Cter tails, respectively; the Nter extension is shown in orange, C) the canonical protein kinase fold represented by the structure of cAMP-dependent protein kinase (PKA) (PDB 1cdk), and D) illustrative examples of the different folds populated by DRDs: the helical subdomain of serine carboxypeptidase Y in green (PDB 1cpy), the human neutrophil defensin 3 in cyan (PDB 1dfn), the protease inhibitor domain of the amyloid beta A4 precursor protein in orange (PDB 1aap) and the somatomedin-B domains of human vitronectin in marine (PDB 1s4g).

Kinases also constitute an interesting model for the analysis of protein aggregation since they appear to have retained a substantial aggregational load, in spite of their uttermost relevance for the maintenance of the cell physiology. Unrelated kinase domains have been shown to form amyloid-like fibrilar structures *in vitro* from substantially unfolded states (Damaschun et al. 2000) or by the increased population of intermediate states caused by mutation (Georgescauld et al. 2011). Even more remarkably, it has been estimated that around 60% of human kinases are clients of Hsp90 (Taipale et al. 2012), which is in good agreement with the frequent observation that kinase domains tend to accumulate as insoluble aggregates when recombinantly overexpressed in bacteria (Benetti et al. 1998; Marin et al. 2010), since prokaryotic Hsp90 appears to lack the ability to chaperone kinases (Buchner 2010).

3.1.b.II.- Ensemble of Disulfide-rich Domains (DRDs)

Many small proteins lack stable hydrophobic cores or secondary structure elements to maintain their native globular structure. Thus supplemental stabilization is achieved in these proteins through binding of metal ions or by the presence of disulfide bonds. Particularly, disulfide crosslinking has a great impact on the stability of the native conformation of proteins, already introduced in section 1.2 and further discussed in the following chapters, and proteins which depend on the presence of disulfide bonds to preserve their native structure usually present more than one disulfide. These proteins may be present as monomeric polypeptides or as components of multimodular proteinaceous entities. Consequently, these proteins have been termed disulfide-rich domains (DRDs), which are specifically defined as small polypeptides with a size usually below 100 aminoacids, that lack an extensive hydrophobic core and stable secondary structure, and whose structure is essentially stabilized by the presence of two or more disulfide bonds in close spatial vicinity (Cheek et al. 2006). DRDs are involved in a wide variety of functions that can be grouped into three major functional classes, including structural roles, enzymatic activity and cellular communication. The latter is perhaps the most widespread function within DRDs, enclosing proteins that function as growth factors, pheromones, enzyme inhibitors and ligand-binding domains.

The particular features of DRDs have made their structural classification at problematic. On one side, the methods of structural classification employed to globally categorize proteins, such as those implemented in the SCOP database (Hubbard et al. 1997) usually rely on a quite strict coincidence of the arrangement of major secondary structure elements, that is not generally applicable to the search of structural homology within DRDs, since disulfide bonding allows for a greater distortion and local disruption of secondary structure elements. On the other hand, specific methods have attempted to classify DRDs based on the spatial super imposition of their disulfide cross-links (Mas et al. 1998) or by defining a “ disulfide signature” according to the disulfide connectivity of the polypeptide, and the length of the “loops” of sequence between consecutive Cys (Gupta et al. 2004; van Vlijmen et al. 2004). However, these latter approaches dismiss important information encoded in sequential and structural similarity, thus questioning the significance of the evolutionary relationships within the categories established according to these methods. In order to overcome these limitations inherent to previous classification methods, Cheek and co-workers underwent a thorough comprehensive classification of DRDs by establishing a hierarchical scheme according to which DRDs are first grouped into “folds”, on the basis of the comparison of structural similarity following a more flexible topological criterion that allows the detection of structural homology between similar arrangements of secondary structure. DRDs are further classified into families according to evolutionary relationships derived from structural, sequential and functional similarities. In this way, 41 “fold groups” have been established for DRDs, comprising 98 families that are globally equivalent to SCOP superfamilies (Figure 12.D).

The nature of disulfide cross-linking allows for a precise characterization of the species populated during the oxidative folding, and the pathway of several DRDs has been described in details (Arolas et al. 2006; Chang 2008), also for the case of structurally equivalent DRDs folding through different oxidative pathways like BPTI (Weissman & Kim 1991) and TAP (Chang 1996). Although the folding pathways of DRDs are among the most well characterized of all globular proteins, the influence of the conformational folding on the variables that determine disulfide formation have prevented to forecast the oxidative folding pathway a given DRD will follow (Arolas & Ventura 2011). The analysis of the reductive unfolding of disulfide-rich proteins has allowed to estimate the overall kind of pathway, either BPTI-like or hirudin-like, a DRD will fold through (Chang 2011); but predicting the oxidative folding pathway from the primary sequence, or even from the knowledge of the native conformation and disulfide pairings, has remained elusive.

In this work we have employed the DRD classification established by Cheek et al, and particularly a dataset of structures representative of DRDs families, in order to evaluate the aggregational and folding determinants of disulfide-rich proteins, which allows to provide new insights on how the interplay between these properties shapes their structures and influences their folding pathways.

3.1.c.- Protein Datasets and Databases

3.1.c.I.- Abundance and Solubility of the *Escherichia coli* Proteome

Escherichia coli is one of the most extensively studied microorganisms and it has been largely exploited in the biochemistry, molecular biology and biotechnology areas as a model organism. Particularly, this bacteria has been employed to unravel key cellular processes such as the functioning of the molecular machinery responsible for proteins synthesis and the assistance of correct folding (Sabate et al. 2010). The major achievement of the development of recombinant DNA technology (Cohen et al. 1973; Morrow et al. 1974), allowed to exploit *E. coli* as a host for the heterologous production of proteins, because of its convenience as an expression system due to the broad knowledge about its physiology, together with its short generation time, ease of handling and growth, and the availability of non-pathogenic strains (Schmidt 2004). However, the recombinant over expression of proteins may lead to the saturation of the bacterial PQC, resulting in the formation of insoluble protein deposits generally termed inclusion bodies (IBs). This aggregated inclusions are mostly composed of the heterologously expressed protein, which may constitute above 90% of the protein content in IBs (Carrió et al. 1998). For a long time, IBs were considered as inactive intracellular “dust balls” of misfolded or unstructured proteins due to their amorphous macroscopic appearance; however accumulating evidence showed later that IBs possess amyloid-like structure (de Groot et al. 2009). This has attracted attention to the use of E. coli as a model system for human depositional disorders (García-Fruitós et al. 2011) like for the study of the role of the PQC on the management of misfolded polypeptide species (Sabate et al. 2010) or the relevance of protein sequence for the tendency to aggregate *in vivo* (de Groot et al. 2006).

Considering the utility of *E. coli* as a model system for the analysis of protein aggregation, we have exploited a series of databases of experimentally determined data about the *E. coli* proteome and transcriptome, in order to analyze the aggregational properties within this system at a proteome-wide level. In this sense, we have integrated experimental solubility data of *E. coli* proteins determined employing an in vitro reconstituted system (Niwa et al. 2009), with different datasets of either abundance of proteins from different subcellular localizations measured employing the APEX mass spectrometry approach, with their associated mRNA levels defined as an average of different techniques (Lu et al. 2007); or cytosolic protein abundance determined using the emPAI mass spectrometry method (Ishihama et al. 2008) with their corresponding mRNA levels obtained by fluorescent DNA microarrays (Bernstein et al. 2002).

3.2.- Bioinformatic Approaches to Predict Protein Aggregation

The vast knowledge accumulated about the determinants of protein aggregation has allowed to develop a variety of mathematical methods aimed for the prediction of protein aggregation in silico. More than 20 computational tools have been published to date, which can be classified into two greater families depending on the nature of the information considered to implement the method of prediction. On one side, **empirical or phenomenological approaches** are based on the identification of experimental variables known to affect the tendency of a polypeptide to aggregate. A different class are the **structure-based methods** which rely on the analysis of the conformational compatibility of polypeptide stretches to the cross-β structural core of amyloid fibrils. The methods may also differ in the type of output they provide, although it commonly comprises the identification of APRs along the polypeptide sequence and its aggregative potential, as well as the relative or virtually absolute tendency to aggregate of the protein. Some predictors provide additional valuable estimates such as the nature of the pairing of the β-strands, either parallel or antiparallel, forming the cross-β core.

In this section a brief description is provided for those methods which have reached a greater spread within the field.

3.2.a.- Phenomenological Methods

The first mathematical tool developed with the aim of predicting protein aggregation was an empirical equation derived from experimental data of the aggregation kinetics of different human muscle acylphosphatase mutants (Chiti et al. 2003). This equation allows to calculate the change in the rate of aggregation into amyloid-like structures upon mutation for unstructured proteins or peptides, on the basis of the change in hydrophobicity, in the propensity to convert from α-helical to β-sheet conformation and in the net charge of the polypeptide caused by the mutation. A further refinement of this equation lead to the development of an algorithm that allows to calculate fibril elongation rates from fully or partially unfolded conformations by considering seven variables, including intrinsic parameters such as hydrophobicity, net charge of the polypeptide, and presence of alternating patterns of hydrophobic and hydrophilic residues, as well as, extrinsic factors like the pH, ionic strength and polypeptide concentration

(DuBay et al. 2004). This algorithm was later adapted to estimate the intrinsic aggregation propensity of the 20 naturally occurring amino acids. In this way, it was possible to calculate the aggregation propensity within a polypeptide sequence in a position-specific manner, where each residue is assigned the average intrinsic aggregation propensity of a window of amino acids centred on it (Pawar et al. 2005). This value of aggregation propensity for each residue is standardized relative to that computed for random sequences with the same length than the sequence analyzed and with the amino acid frequencies of the Swiss-Prot database. Therefore, this modification of the method by Dubay et al. allows to detect APRs as those fragments of the sequence with consecutive residues possessing an intrinsic aggregation propensity above 1 standard deviation.

Following a similar rationale to that of the first mathematical expression, Tartaglia and co-workers developed a function to calculate the change in the aggregation rate upon mutation that, aside from the change in the propensity to β-conformation and in charge, also takes into account changes in the accessible surface area, the number of aromatic residues, which influence the extent of π-stacking, and the dipolar moment of polar sidechains (Tartaglia et al. 2004). An advantage of this method was the absence of free parameters, thus allowing a broader generalization for the prediction of aggregation. This method was also modified later in order to allow for the prediction of absolute aggregation rates and the detection of APRs, further estimating the preferred orientation, either parallel or antiparallel, of the β-aggregating segments (Tartaglia et al. 2005).

Zyggregator is the latest development of the concept introduced by Chiti and co-workers in order to predict protein aggregation. It adds to the implementation by Pawar et al. a parameter in order to account for the impact of gatekeeper residues against aggregation and, in contrast to the previous developments that only allowed to predict aggregation from fully or partially unfolded states, it also incorporates the influence of local structural stability (Tartaglia et al. 2008; Tartaglia & Vendruscolo 2008); on the basis of the prediction of the flexibility and solvent accessibility of the polypeptide chain as implemented in the CamP method (Tartaglia et al. 2007). Therefore, this algorithm allows to predict the aggregation propensity in structured proteins as well.

The first method that allowed the evaluation of the tendency of a protein to aggregate from its plain sequence was the TANGO algorithm (Fernandez-Escamilla et al. 2004). This algorithm is based on a statistical mechanics concept where several states are defined including the random coil and native conformations, and α-helix, β-turn and β-sheet aggregate states, which are characterized by physical-chemical properties, as well as by conformational and statistic compositional preferences. TANGO calculates the population by fragments of the sequence of each state, according to a partition function, where the population of a particular state is proportional to the energy of the fragment on it. This energy is obtained taking also into account physical-chemical variables such as the pH, temperature and ionic strength, or extrinsic factors like trifluoroethanol (TFE) concentration. TANGO predicts a segment of the sequence to possess tendency to aggregate when its length is of at least 5 residues and it populates the β-sheet aggregate state with a probability higher than 5%. In this way, TANGO was also the first method to allow for the detection of APRs within protein sequences.

On the other hand, AGGRESCAN was the first predictor of protein aggregation which was specifically based on empirical data obtained *in vivo*. By exploiting a strategy developed to determine protein aggregation in bacterial cells employing GFP (Waldo & Standish 1999), a library of point mutants of the Aβ42 peptide fused to the GFP was constructed, which allowed to set up a scale of intrinsic aggregation propensity *in vivo* for the 20 naturally occurring amino acids (de Groot et al. 2006). AGGRESCAN predicts aggregation propensity from the primary sequence by computing for each amino acid the average aggregation propensity of a variable window, depending on the size of the protein, centered at this position (Conchillo-Solé et al. 2007). An APR or "hot-spot" is defined whenever a stretch of 5 or more consecutive residues is detected within the polypeptide whose computed aggregation propensities are above a defined threshold, which corresponds to the average value of the aggregation propensity scale. The AGGRESCAN algorithm has been implemented as a web server and constitutes an extremely versatile tool, allowing to calculate the aggregation properties of either single polypeptides or large protein ensembles, and providing multiple parameters to establish comparisons between individual proteins or datasets, such as size-normalized absolute aggregation propensities and number of APRs per molecule, and indicators of the aggregative potency of the detected APRs.

Finally, the Simple ALgorithm for Sliding Averages (SALSA) assumes a strong correlation between the propensity to adopt β-strand conformation and the ability to form fibrillar structures. Therefore, it aims to identify APRs as regions of the polypeptide sequence with strong β-strand propensity. It does so by assigning to each residue the mean β-strand propensity of different averaging windows centered on it, according to the secondary structure propensities defined by Chou and Fasman (Zibaee et al. 2007).

3.2.b.- Structure-based Methods

The structure-based approaches rely on specific structural features associated to the formation of ordered amyloid-like aggregates. The first method of this kind was the Net-CSSP algorithm which exploits the concept of "chameleon sequences" or "conformational switches" as relevant conformational transitions in the formation of amyloid-like structure. This approach is based in the observation that the regular secondary structure adopted by a polypeptide is dependent on its tertiary contacts (Minor & Kim 1996), in such a way that certain sequences without β-conformation may encode, in the context of their structural environment, a hidden β-propensity (Yoon & Welsh 2004). In this sense, Yoon n Welsh have developed the Contact-dependent Secondary Structure Prediction (CSSP) algorithm in order to detect sequence stretches with noticeable hidden β-propensity, which could act as potential “conformational switches” (Yoon & Welsh 2004; Yoon et al. 2007).

A different concept is exploited by FoldAmyloid, this method is based on the consideration that the packing density of amino acids allows to describe the conformational properties of a polypeptide structure, in such a way that amyloid-like conformations correspond to those exhibiting a higher packing density. By inspecting the SCOP structural database, a mean packing density scale was defined for the 20 naturally occurring amino acids, on the basis of the average number of contacts, defined according to spatial criteria, observed for each residue (Garbuzynskiy et al. 2010). An amyloidogenic APR is detected according to this method when more than 5 consecutive residues with a mean packing density above a certain threshold are found.

In contrast to these methods, a majority of structure-based approaches are based on the analysis of the specific features of the β-sheet structure in the cross-β core of amyloid-like aggregates. One of such methods is the Prediction of Amyloid Structure Aggregation (PASTA) algorithm, which relies on the idea that β-conformation in amyloid structure corresponds to pairings of in-register β-sheets, either parallel or antiparallel, with a minimum energy. PASTA calculates pairing energies for a intermolecular β-sheet on the basis of individual pairing energies between each two amino acids facing each other in the β-sheet; such individual pairing energies have been statistically derived from the observed distribution of each pair of amino acids found in β-sheet conformation, either parallel or antiparallel, within a refined dataset of non-redundant crystal structures of globular proteins with diverse folds (Trovato et al. 2006). β-sheet energies are calculated by computing intermolecular pairings, both in a parallel or antiparallel fashion, between identical continuous stretches of variable length within the input sequence, while the rest of the polypeptide is considered disordered. The total pairing energies result from the sum of the individual pairing energies of each couple of contacting residues, after introducing a correction for the loss of entropy derived from the ordering of the residues within the stretches. Those pairings of stretches with energies below a certain threshold are considered to present an increased likelihood to form the cross-β core in amyloid-like structures and are identified as APRs. In contrast with other prediction tools, PASTA predicts, in addition, the preferred pairing orientation, whether parallel or antiparallel, an APR would adopt in that cross-β core.

A similar approach is employed by BETASCAN, which computes the most probable β-strand pairing between pairs of segments of the polypeptide chain with the same length. The likelyhood of the pairings is calculated according to a table of probability for couples of residues to be H-bonded in amphiphilic β-sheets of structures selected from the PDB (Bryan et al. 2009).

The Peptide Interaction Matrix Analyzer (PIMA) method is based in the β-pairing concept as well. In this case, however, the pairings are not defined on the basis of statistically derived information, but it calculates the energy of every possible peptide segment threaded onto a in-register, parallel or antiparallel, β-sheet structure by employing a physics-based forcefield (Bui et al. 2008).

Another type of structure-based methods are those which rely on the properties of the solved three-dimensional structures of small peptides forming amyloid-like structures. The first within this class of approaches was the 3D profile method, it works by threading every hexapeptidic segment of a polypeptide sequence not containing a Pro onto the backbone of the three-dimensional structure derived from a crystal of the fibril-forming peptide NNQQNY (Thompson et al. 2006). The fit of each hexapeptidic fragment to this template structure is evaluated by energy minimization with the RosettaDesing program (Kuhlman & Baker 2000). Those segments yielding an energy value below a defined threshold are predicted to have a high propensity to form amyloid-like structures.

Waltz is the latest prediction algorithm that takes advantage of the information obtained from small fibril forming peptides. This method is specifically intended to detect amyloidogenic stretches, not mere β-aggregation propensity, and employs a position-specific scoring matrix (PSSM) derived from the sequences of 48 fibril-forming hexapeptides. Waltz identifies amyloidogenic fragments in polypeptide sequences according to a scoring function composed of a first sequential parameter which measures the suitability of the sequence according to the Waltz PSSM, a second physical-chemical parameter that weights a series of relevant variables for amyloid formation, and a third structural parameter that evaluates the conformational fitting of the sequence stretch to the structural template of the GNNQQNY fibril forming peptide employing another position-specific pseudoenergy matrix (Maurer-Stroh et al. 2010).

3.2.c.- Consensus Methods

Each of the methods described above develops one or several specific determinants that are considered as relevant for the aggregation of polypeptides into ordered structures. However, none of them incorporates the whole ensemble of different concepts exploited, which might be complementary for the detection of the deposition of polypeptides into β-sheet enriched aggregates. The idea that combining the outputs provided by different algorithms might increase the accuracy of the prediction by improving the sensitivity towards the divergent strength of the several variables influencing different modes of aggregation, and at the same time by minimizing the method-specific bias towards the overprediction of certain types of aggregation, has been exploited in the AMYLPRED server. It builds a consensus prediction at a residue level by integrating the results given by different previously published methods. The initial version incorporated 5 methods to construct the consensus prediction (Frousios et al. 2009), but a recent release can include up to 11 different methods (Tsolis et al. 2013).

3.2.d.- Assessment of the Predictive Value of Algorithms

Most of the algorithms described above have been developed, and in some cases parametrized, on the basis of properties of amyloid-like aggregates derived from the *in vitro* experimental characterization of the reaction of aggregation of certain model proteins. AGGRESCAN is the sole exception, since it was developed from a experimentally derived scale of intrinsic aggregation propensity for the natural amino acids.

Since the complex cellular environment influences strongly polypeptide deposition, and indeed possesses mechanisms to control it, the question arises as to whether these prediction tools are able to predict protein aggregation *in vivo*. In order to assess this issue, Chiti and co-workers have evaluated the ability of different algorithms to predict the depositional properties of polypeptides *in vivo*, by employing several datasets of proteins whose tendency to aggregate had been determined experimentally (Belli et al. 2011). In general terms, the predictors are able to forecast protein aggregation *in vivo* with significant accuracy, being phenomenological approaches the ones that perform globally better than structure-based methods. This might indicate that phenomenological methods allow to capture the determinants of *in vivo* aggregation more precisely than structure-based ones. The general predictive capability of the algorithms is, anyway, in good agreement with similar trends observed when analyzing large ensembles of proteins, using different methods.

Interestingly, AGGRESCAN is the algorithm that yields a better global performance across the different datasets analyzed. This does not come as a surprise considering this algorithm was developed employing an aggregation scale derived from experiments *in vivo*. Nonetheless, the good performance of AGGRESCAN, together with the different sources of the proteins composing the ensembles used to test the algorithms, provides another piece of evidence for the suitability of *E. coli* as a general model system for the analysis of protein aggregation.

CHAPTER 4.- EVOLUTIONARY STRATEGIES TO PREVENT PROTEIN AGGREGATION

4.1.- Introduction

The ability to aggregate into a cross-β enriched conformation seems to constitute a generic property of polypeptides,. A closely related feature, the presence of at least one APR in most proteins of different proteomes has also been observed employing the power of computational tools for the prediction of protein aggregation (Rousseau et al. 2006; Conchillo-Solé et al. 2007). Taking into account the deleterious consequences protein aggregation may exert on cell homeostasis it expected that natural selection has acted to contra balance this effect and minimize the aggregation propensity of natural polypeptides (Monsellier & Chiti 2007; Sanchez de Groot et al. 2012).

Indeed, the PQC described in section 1.5.b can be regarded as an ensemble of mechanisms cells have evolved to control and manage the formation of misfolded species. Nonetheless, evolutive pressure has also acted on protein sequences in order to minimize their tendency to aggregate, as exemplified by correlations observed between ascending organism complexity and decreasing aggregation propensity of proteomes, independently of the method employed to predict protein aggregation., (Tartaglia et al. 2005; Rousseau et al. 2006),.

In first place, the ability to efficiently attain a compact globular structure, commonly referred to as foldability, has been regarded as a side strategy evolution has exploited to avoid aggregation (Monsellier & Chiti 2007). Since interactions of hydrophobic nature are considered to possess a dominant role in the initial stages of both folding (Kauzmann 1959) and aggregation (Cheon et al. 2007; Auer et al. 2008), these are regarded as competing processes. However, while protein folding is guided by the establishment of intramolecular interactions, aggregation is dominated by the formation of intermolecular contacts. Therefore, protecting APRs by substantially burying them in the hydrophobic core of stable globular structures serves as a mechanism to avoid the spurious interactions leading to aggregation. Consequently, the more stable the native conformation of a protein, the less likely it will populate partially unfolded states allowing the establishment of aberrant intermolecular contacts. In fact, destabilization of the native conformation by mutations has been shown to promote the formation of different types of aggregates of proteins, independently of them being associated to pathologies (Chan et al. 1996; Wall et al. 1999) or without any known linkage (Chiti et al. 2000; Espargaró et al. 2008; Castillo et al. 2010). A series of evidences suggest evolutionary pressure acts to select for more stable structures within the constraints imposed by functionality. First, theoretical studies of protein folding based on polymer statistical mechanics have shown that globular proteins have evolved to fold rapidly and cooperatively to the native state, because their energy landscapes are “minimally frustrated” (Wolynes 2008), which implies that partially folded conformations tend to be short-lived and also provide polypeptides with robustness against sequential change, in the sense that most of the possible single point substitutions would have little effect on stability. Accordingly, the folding of proteins lacking an “evolutive history” has been experimentally proved to be poorly cooperative (Watters et al. 2007). As well, the analysis of mutational effects on globular proteins has shown that the major evolutionary constrain, in the absence of functional restrictions, is protein stability and, interestingly, an increasing tendency to aggregate may be tolerated as long as stability is preserved (Sánchez et al. 2006). Along the same line, it has been suggested that evolution tends to shape proteins to fold faster (Debès et al. 2013), although the expansion of fold diversity makes difficult to track this effect along the evolutionary history. The relevance of an efficient folding as a side strategy against aggregation is further highlighted by the observation that longer proteins, which are theoretically predicted to fold slower as protein size increases (Ivankov et al. 2003), tend to present less potent APRs (Monsellier et al. 2008).

The aforementioned selection for a stable native state relies mainly on the analysis of polypeptide conformations at the level of tertiary structure. Nonetheless, several proteins with a stable oligomeric quaternary structure are known to experience a pathogenic conversion to form amyloid structures, including transthyretin (TTR), superoxide dismutase 1 (SOD1) and β2-microglobulin (β2m). The statistical analysis of these and other natively oligomeric proteins and complexes has revealed that the physical-chemical properties of the amino acids involved in the establishment of protein-protein interfaces are of the same nature than those of the residues enriched in APRs (Pechmann et al. 2009; Castillo & Ventura 2009), in such a way that interaction interfaces and APRs overlap significantly. Therefore, the formation of stable quaternary structures and protein complexes could exert a protective role by shielding APRs, avoiding the establishment of intermolecular interactions (Masino et al. 2011). Moreover, this shielding character appears to be fostered by the presence of disulfide bonds and attractive electrostatic interactions in the proximity of interfaces, enhancing both their stability and specificity (Pechmann et al. 2009).

Nature also seems to have shaped proteins to avoid aggregation at the secondary structure level, by incorporating negative design strategies in the structures of all-β proteins (Richardson & Richardson 2002). The peripheral strands flanking β-sheets possess one side which is free to establish hydrogen bonds with neighboring molecules, a feature increases the risk to establish non-functional intermolecular contacts. A straightforward means to prevent such a risk is to avoid the presence of edge strands themselves by forming continuous β-strands, as in β-barrel structures. As well, in non-continuous β-sheets, peripheral strands are usually capped with loops or helices, or may be present as short twisted strands that disrupt the geometry of the sheet at its ends. Partial β-strand distortion can also be achieved by the introduction of β-bulges, charges pointing “inward” towards adjacent backbone elements or by inducing bends in the strand with the incorporation of Pro or Gly residues.

The mutational analysis of aggregation-prone proteins, and the statistical survey of the compositional features of APRs and its environment have allowed to establish the determinants of protein aggregation, that have already been discussed in section 1.4.b. These determinants reveal, how evolution has shaped proteins primary sequences to reduce their propensity to aggregate. As it has already been mentioned, patterns of alternating polar and non polar residues are underrepresented in protein sequences relative to any other possible amino acid combinations (Broome & Hecht 2000). These alternating patterns are considered to favor the formation of amphiphilic β-sheets, therefore increasing the likelihood to aggregate into amyloid-like structures. Their underrepresentation suggests this patterns are evolutionarily disfavored. In a similar way, it has been shown that continuous stretches of three or more hydrophobic residues are less frequent than it would expected (Schwartz et al. 2001) and that the polar content of buried blocks of residues tends to increase with the size of the block (Patki et al. 2006). Altogether, these findings point to a selective pressure against the presence of long continuous segments of hydrophobic amino acids since, because their physical-chemical properties promote the emergence of APRs within polypeptide sequences. Another feature which is regarded as an evolutionary strategy exploited in the primary structure of polypeptides to confront aggregation propensity, is the recurrent presence of gatekeeper residues flanking APRs (Rousseau et al. 2006). Likewise, although as discussed previously the effect of point substitutions appears to act preferentially on the preservation of protein stability rather than on the avoidance of the tendency to aggregate, a few examples suggest that the observed conservation of the b-breaker amino acids Pro and Gly in certain protein domains (Steward et al. 2002; Parrini et al. 2005) is the result of a selective pressure to avoid aggregation.

The early observation that protein deposits associated to different pathologies are mostly enriched in a single polypeptide sequence (Dobson 2001), together with the studies showing the impact of sequential divergence on the seeding of protein aggregation kinetics (Krebs et al. 2004; Sabaté et al. 2010) indicate the relevance of sequence identity in the protein aggregation process. This feature puts multimodular proteins composed of successive homologous domains at a high risk of aggregation due to the increased local effective concentration inherent to this kind of proteins. Not surprisingly, it has been shown that adjacent domains in multimodular proteins tend to possess a lower sequence identity than non-adjacent ones (Wright et al. 2005), so it has been argued that increased sequential divergence between adjacent domains would constitute yet another evolutionary strategy to decrease the risk of aggregation. Homooligomeric proteins are another kind of polypeptides which present a inherently high local monomer concentration and, despite the formation of protein-protein functional interfaces can be regarded, as described before, as a protective mechanism against aggregation, in their case the impact of sequence identity promoting aggregation reaches its maximum. Nonetheless, it has been found that proteins with a native homooligomeric quaternary structure generally present a lower aggregation propensity relative to those that do not adopt a functional homooligomeric conformation (Chen & Dokholyan 2008); thus indicating that proteins which are inherently at a higher aggregation risk, also experience a greater pressure to minimize their tendency to aggregate.

Since, as discussed before, folding into a stable globular conformation exerts a protective role against protein aggregation, the question arises about how proteins lacking a defined three-dimensional structure in its physiological environment, such as IDPs, avoid the risk to adopt aberrant conformations and aggregate. It has already been introduced that IDPs possess specific sequential features which differentiate them from globular proteins (Uversky 2002), providing them with a higher net charge and reduced hydrophobicity. They are particularly depleted in aggregation-prone residues and, consequently, exhibit a significant lower content of APRs compared to globular proteins (Linding et al. 2004). Moreover, they are enriched in charged and β-breaker residues such as Pro (Tompa 2002). These properties make IPDs and IDPRs more resistant to protein aggregation, and it has been proposed that IDPRs in the N- and C-terminal regions of globular proteins can act as entropic bristles generating an excluded volume around protein molecules (Uversky 2013b), thus protecting the globular domain from the establishment of spurious intermolecular contacts. Computational models do, in fact, illustrate how IDPRs, either terminal (Abeln & Frenkel 2008) or internal (De Simone et al. 2012), generate such an entropic hindrance that prevents intermolecular interactions, subsequently abrogating aggregation. These findings suggest that enrichment in IDPRs within globular proteins may also serve as a mechanism to prevent protein aggregation.

Finally, the high order nature of the aggregation reaction makes it deeply dependent on the local protein concentration. Several findings suggest different strategies have been evolved in order to tightly regulate protein populations within the cell. First, a correlation has been observed between mRNA levels and predicted aggregation propensity in the human proteome (Tartaglia & Vendruscolo 2009), indicating that protein expression is finely tuned depending on the aggregation properties of the polypeptides. Interestingly, the structural resolution of the polysome assembly in the simultaneous translation of single mRNA molecules by multiple ribosomes has revealed an helical arrangement (Brandt et al. 2009). Such an organization has been considered to allow for the emergence of nascent polypeptides in different directions, reducing its effective local concentration and, consequently, lowering the risk of co-translational aggregation. Protein population can also be controlled at the level of the degradation machinery, in this sense, an evolutionary trend has been revealed by an study that analyzed the aggregation propensity of *E. coli* proteins relative to their turnover rates (De Baets et al. 2011). It was observed that long-lived proteins possess a lower tendency to aggregate, while a higher aggregation propensity may be preserved in those proteins that are degraded fast.

In the following sections, a more specific analysis is presented for three particular aspects of the aforementioned evolutionary strategies to prevent aggregation: the impact of disulfide bonding on protein stability and its association to the aggregation of polypeptides, the effect of IDPRs enrichment at the terminal regions of globular proteins, and the role of protein abundance.

4.2.- Impact of Disulfide Cross-linking on the Formation and Toxicity of Amyloid Fibrils

Folding into a stable globular structure is regarded, as mentioned above, as a primary protection mechanism against the establishment of aberrant intermolecular contacts. On the other hand, it has also been introduced in section 1.2 that disulfide bonds provide a great stabilization to the native state of globular proteins, which is mostly attributed, based on the theoretical analysis of polymeric chains (Poland & Scheraga 1965), to a destabilization of the denatured state caused by a decrease on its configurational entropy.

Many proteins associated with pathological aggregation into amyloid-like fibrils present disulfide bonds (Mossuto 2013). For example, mutations linked to familial Amyotrophic Lateral Sclerosis (ALS) cause SOD1 intramolecular disulfide bond to be more susceptible towards reduction, leading to a destabilization of the mature protein and increasing the population of unfolded states (Tiwari & Hayward 2003; Furukawa & O'Halloran 2005; Kayatekin et al. 2010). Also, the formation of an intramolecular disulfide bond has been proposed to decrease the rate of fibrillation for certain Tau isoforms (Barghorn & Mandelkow 2002). However, the precise role of disulfide bonding in proteins involved depositional disorders is still poorly understood.

In order to analyze the specific impact of disulfide cross-linking on the process of amyloid fibril formation, we have employed the PI3-SH3 domain as a model. The aggregation into amyloid-like fibrils of this SH3 domain, which is not associated to any known depositional disorder, has been extensively characterized (Guijarro, Morton, et al. 1998; Ventura et al. 2002). PI3-SH3 is able to form fibrils departing from the "A-state" that this domain populates under acidic conditions, and whose compaction is lower than that of the native conformation, but greater than the observed in the denatured state.

Engineering a disulfide that cross-links the N and C termini of PI3-SH3 does not alter the conformation of the native state, neither that of the A-state. This disulfide bond largely enhances, as expected, both the thermal and chemical stability of the domain, yielding one of the greatest stabilizations reported for an SH3 domain so far. However, the resulting stability increase is not as large as it would be theoretically derived from the decrease in the configurational entropy of its unfolded state. This suggests that, although it may constitute the dominant effect contributing to stability changes upon disulfide bonding, restriction of the configurational entropy is not the only factor that determines the actual increment in stability. This rather arises from the balance between different entropic and enthalpic components (Doig & Williams 1991; Betz 1993), as it has been observed for several proteins (Betz & Pielak 1992; Johnson et al. 1997; Hagihara et al. 2007). As a result of the decrease in the configurational entropy of the unfolded state, PI3-SH3 folds faster to its native state, while its unfolding rate decreases only slightly.

The stabilization of the native state induced by the engineered disulfide bond does not impede the population of the A-state by the SH3 domain and, subsequently, it does not prevent the formation of amyloid-like fibrils. Nonetheless, cross-linking strongly influences the aggregation kinetics of PI3-SH3, which has been shown consistent with a NCC mechanism, by decreasing both its nucleation and elongation rates. Since the disulfide bond does not appear to affect the conformation of the SH3 domain in the A-state, the reduction on its nucleation rate likely arises from the restrictions imposed by the cross-link on the conformational flexibility that is required for the polypeptide to convert into a conformation compatible with the cross-β spine of the fibril. The lower elongation rate also implies that such restriction on the conformational flexibility does not only affect the conversion into structures enriched in intermolecular β conformation but also constraints the docking of oligomeric subunits in the elongation of the fibril.

The different aggregation kinetics is also translated into different morphological and conformational properties of the fibrils formed by the SH3 domain with an engineered disulfide. Its amyloid-like fibrils present a significantly shorter length, lower exposure of hydrophobic patches, increased content of intermolecular β-sheet and higher chemical stability than the fibrils formed by the wt protein. The shorter length might be a reflection of the flexibility constraints imposed by the disulfide bond, while increased stability of the fibrils may also contribute to the decrease of the nucleation and elongation rates, since a greater stability implies a lower population of mature oligomers by fibril depolymerization, thus dimishing the contribution of secondary pathways to nucleation (Knowles et al. 2009). The structural model proposed for the amyloid-like fibrils of wt PI3-SH3 based on MAS-ssNMR (Bayro et al. 2010) provides insightful clues to interpret the higher stability observed for the fibrils of the engineered domain; while in the wt fibrils the Cter of the polypeptide exhibits a significant dynamic flexibility at the fibril surface, the engineered disulfide likely constrains this region into the cross-β core the Nter region is forming part of. This is consistent with their higher β content and decreased hydrophobic exposure of the fibrils formed by the engineered domain.

Interestingly enough, the fibrils formed by the engineered domain exhibit a lower toxicity for cultured cells than those of the wt PI3-SH3. This is along the same line of the higher cytotoxicity observed for lysozyme fibrils formed under reducing conditions (Mossuto et al. 2011), which also present a lower content of intermolecular β conformation compared to the fibrils formed by lysozyme under conditions where its disulfides remain intact. The lower toxicity of the fibrils formed by the disulfide-crosslinked proteins in these cases likely results from a larger content of regular structure and, particularly, from a decreased exposure of hydrophobic patches, since such exposure of hydrophobic regions has been shown to correlate with aggregate toxicity (Bolognesi et al. 2010).

These results suggest that the effect of disulfide bonds in the stabilization of the native state, but as well in constraining the kinetics of aggregation or in yielding fibrils with reduced cytotoxicity, may have a protective role against the deleterious effects of protein aggregation. Consistent with this rationale, extracellular SH3 domains, which present two conserved disulfide bridges, exhibit a higher sequential aggregation propensity. In a similar study, Mossuto and coworkers found that human proteins borned with disulfide bonds present a higher aggregation propensity compared to proteins devoid of disulfides, with extracellular polypeptides bearing disulfide bonds being the ensemble with the greater aggregation propensity. Together, these considerations indicate the impact of disulfide cross-linking may serve as a strategy to tolerate a higher aggregational load, particularly for proteins that develop their functions in harsh environments for their native states, such as the extracellular space.

ANTIOXIDANTS & REDOX SIGNALING
Volume 16, Number 1, 2012
© Mary Ann Liebert, Inc.
DOI: 10.1089/ars.2011.3936

ORIGINAL RESEARCH COMMUNICATION

Contribution of Disulfide Bonds to Stability, Folding, and Amyloid Fibril Formation: The PI3-SH3 Domain Case

Ricardo Graña-Montes,[1,*] Natalia S. de Groot,[1,*] Virginia Castillo,[1] Javier Sancho,[2–4]
Adrian Velazquez-Campoy,[2–5] and Salvador Ventura[1]

Abstract

Aims: The failure of proteins to fold or to remain folded very often leads to their deposition into amyloid fibrils and is the origin of a variety of human diseases. Accordingly, mutations that destabilize the native conformation are associated with pathological phenotypes in several protein models. Protein backbone cyclization by disulfide bond crosslinking strongly reduces the entropy of the unfolded state and, usually, increases protein stability. The effect of protein cyclization on the thermodynamic and kinetics of folding has been extensively studied, but little is know on its effect on aggregation reactions. ***Results:*** The SRC homology 3 domain (SH3) of p85α subunit of bovine phosphatidyl-inositol-3′-kinase (PI3-SH3) domain is a small globular protein, whose folding and amyloid properties are well characterized. Here we describe the effect of polypeptide backbone cyclization on both processes. ***Innovation:*** We show that a cyclized PI3-SH3 variant is more stable, folds faster, aggregates slower, and forms conformationally and functionally different amyloid fibrils than the wild-type domain. ***Conclusion*:** Disulfide bridges may act as key molecular determinants of both productive protein folding and deleterious aggregation reactions. *Antioxid. Redox Signal*. 16, 1–15.

Introduction

Protein misfolding and aggregation is associated to an increasing number of human disorders, ranging from dementia to diabetes (13). The proteins involved in such pathologies are not related sequentially or structurally, but self-assemble into ordered amyloid fibrils sharing a common cross-β-sheet motif (23). This property is not an unusual feature exhibited by a reduced set of proteins, but is rather inherent to most, if not all, polypeptides (22, 28). Thus, in addition to the native fold, an alternative, stable, and ordered state is accessible to proteins. Both states compete in the cell, resulting in functional or toxic conformations depending on whether the native or the aggregated state is populated (30).

Because protein aggregation requires at least local backbone fluctuations and in many cases partial unfolding (12, 14), the deposition propensity of globular proteins appears to be linked to their thermodynamic and/or kinetic stability (10), in such a way that mutations or environmental conditions that destabilize the native structure or increase unfolding rates are associated to pathological phenotypes in several protein models (6, 24). This could be one of the underlying reasons why proteins that function in harsh environments, such as the extracellular space, have evolved disulfide bonds (48). Natural disulfide bonds can stabilize proteins to a large extent and strongly reduce conformational fluctuations (1). In some cases, the stabilizing effect is so high that the protein readily unfolds when its disulfides are reduced (19). By crosslinking sequentially distant regions of the polypeptide chain, disulfide bonds are assumed to decrease the entropy of the unfolded ensemble, making it less favorable compared with the folded conformation. Thus, the maximum stabilization will be attained when the disulfide bond links the two termini of the molecule; natural proteins like cyclotides exploit this strategy to become exceptionally stable (9). Crosslinking the polypeptide chain affects not only the stability of the protein at equilibrium but also, in many cases, its mechanism and/or kinetics of folding (25, 49). In fact, engineering of new disulfide bonds and analysis of the corresponding changes in folding kinetics is a useful approach to

[1] Departament de Bioquímica i Biologia Molecular, Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain.
[2] Departamento de Bioquímica y Biología Molecular, Universidad de Zaragoza, Zaragoza, Spain.
[3] Institute of Biocomputation and Physics of Complex Systems, Universidad de Zaragoza, Zaragoza, Spain.
[4] Unidad Asociada BIFI-IQFR, CSIC, Zaragoza, Spain.
[5] Fundación ARAID, Diputación General de Aragón, Zaragoza, Spain.
*These two authors contributed equally to this work.

Innovation

Cyclization significantly stabilizes SH3 domain of p85α subunit of bovine phosphatidyl-inositol-3′-kinase (PI3-SH3) and strongly accelerates its folding rate while having a minor effect on its unfolding rate, promoting the fast attainment of the folded state and a higher proportion of native molecules at equilibrium, contributing thus to decrease the population of aggregation-prone species. In addition, even when amyloidogenic intermediates are populated, cyclization slow downs PI3-SH3 self-assembly and renders the fibrils less cytotoxic. Thus, disulfide bonds likely reduce the chances of pathogenic protein aggregation, especially in the extracellular space, where polypeptides are not subjected to the surveillance of molecular chaperones. Prevention of protein aggregation is an important selective pressure acting on the evolution of protein sequences (11); the protective effect of disulfide bonds against toxic protein deposition in the oxidative extracellular background might explain why the sequences of secreted SH3 domains can support a higher intrinsic aggregation load than those of their less stable intracellular counterparts. Importantly, during the revision of this work, a study by Mossuto *et al*. showed that reduction of disulfides in human lysozyme increases the toxicity of the resulting fibrils and, more generally, that disulfide-containing proteins in the human proteome display a high aggregation propensity (36). Overall, it appears that the impact that covalent polypeptide crosslinking exerts in the complex processes of protein folding and aggregation might be more relevant from the physiological point of view than previously thought. In this sense, the design of new disulfide bonds might be an effective strategy to reduce the aggregation of target proteins of biotechnological interest.

study the proximity of specific protein regions in the transition state ensemble (27).

SRC homology 3 domain (SH3) domains are good model systems for folding and aggregation analysis due to their monomeric state, their small size, and the absence of cofactors and, usually, disulfide bonds. The SH3 domain of p85α subunit of bovine phosphatidyl-inositol-3′-kinase (PI3-SH3) consists of 83 residues that fold into five β-strands and two short helix-like turns. The five β-strands are arranged in two β-sheets that are orthogonal to each other forming a β-sandwich (33) (Fig. 1). PI3-SH3 appears to fold/unfold following a canonical two state mechanism (49), although recent studies suggest the existence of a transient intermediate after the rate-limiting step of folding (47). PI3-SH3 constitutes one of the best-characterized globular proteins able to form ordered amyloid fibrils starting from an acid-unfolded state (28, 46). Therefore, this domain provides a framework to study the effect of backbone cyclization on protein stability, folding, and aggregation.

Here, we investigate the properties of a PI3-SH3 mutational variant in which the N- and C- termini have been crosslinked by a designed disulfide bond. Overall, the data collected here indicate that, by altering polypeptide chain connectivity, disulfide bridges have a high impact in protein folding as well as in amyloid formation and suggest that they may play a major role in minimizing misfolding and aggregation phenomena.

Results

Crosslinking the N and C termini in the PI3-SH3 domain

We used FoldX (42) to design a PI3-SH3 domain with the N and C terminal ends linked by a disulfide bridge. We searched for two residues close enough in the native structure to permit the formation of the desired covalent interaction after their mutation to Cys. The best positions were Ala3 and Ile82, due to their proximity (distance between the β-carbons = 4.2 Å), face-to-face orientation, and proper dihedral angles. Therefore, we cloned, produced, and purified an SH3-PI3 mutant with both residues replaced by Cys (PI3-SH3-SS). The absence of detectable alkylation by vinylpyridine confirmed the formation of the designed disulfide bond. The introduced covalent linkage closes off a loop of 78 residues (Fig. 1).

Conformational analysis of PI3-SH3 domains under native conditions

The wild-type domain (PI3-SH3-WT) and PI3-SH3-SS exhibit analogous far-UV circular dichroism (CD) spectra under native conditions (pH 7.0 and 298 K) with a characteristic minimum at 222 nm (46) (Fig. 2A). However, the PI3-SH3-WT maximum at 235 nm is absent in PI3-SH3-SS. The far-UV CD spectrum of PI3-SH3-SS under reducing conditions does not change significantly, suggesting that the change in the spectrum is associated to the sequential changes and not to disulfide bond promoted strain. Accordingly, Cys residues are known to contribute negatively to the CD signal of peptides in this region of the spectrum (32). We will refer to PI3-SH3-SS in its reduced form as PI3-SH3-$(SH)_2$.

PI3-SH3 contains a single Trp in position 55 and 4 Tyr residues at positions 6, 8, 12, and 59. We excited the different proteins at 268 nm and recorded their intrinsic fluorescence emission spectra under native and reducing conditions. No significant differences were detected between domains (Fig. 2B). The structural features of the PI3-SH3-WT and PI3-SH3-SS were also evaluated by NMR spectroscopy. The one-dimensional NMR (^{1}H-NMR) spectrum of both proteins at pH 7.0 and 298 K display 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. 3). The two spectra are almost identical, which together with the CD and fluorescence data indicate that the introduced mutations do no affect significantly the PI3-SH3-fold.

Thermal unfolding of PI3-SH3 domains

The thermal stability of PI3-SH3 domains at pH 7.0 was analyzed by intrinsic fluorescence, CD, differential scanning calorimetry (DSC), and ^{1}H-NMR (Fig. 4).

The transition curves from heat-induced fluorescence emission changes in PI3-SH3 domains are shown in Figure 4A. The thermal denaturation curves followed by far-UV CD at 235 nm are show in Figure 4B. In both cases a single cooperative transition was observed and the data could be fitted accurately to a two-state temperature-induced unfolding

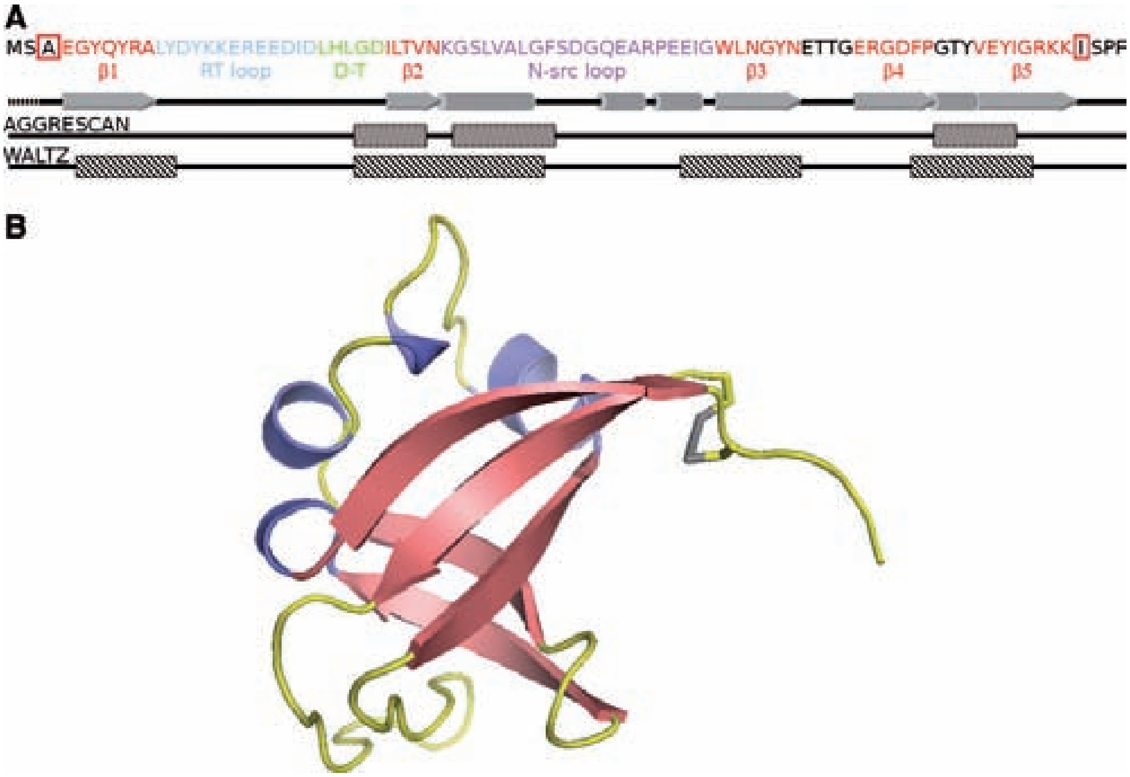

A
MSAEGYQYRALYDYKKEREEDIDLHLGDILTVNKGSLVALGFSDGQEARPEEIGWLNGYNETTGERGDFPGTYVEYIGRKKISPF
β1
RT loop
D-T
β2
N-src loop
β3
β4
β5
AGGRESCAN
WALTZ
B

FIG. 1. **SH3 domain of p85α subunit of bovine phosphatidyl-inositol-3′-kinase (PI3-SH3)-WT. (A)** Sequence, SH3-fold characteristic structural features are depicted with different colors; the positions mutated by Cys are shown within *red squares*. The diagram below represents the distribution of secondary structure elements along the sequence. Aggregation-prone regions of PI3-SH3-WT predicted with AGGRESCAN and WALTZ are shown as wavy and diagonal lines, respectively. **(B)** Structural model of PI3-SH3-SS (1PHT.pdb), sulfur atoms forming the disulfide bond of the mutated PI3-SH3 domain are colored in gray. The model was generated using PyMol. (To see this illustration in color the reader is referred to the web version of this article at www.liebertonline.com/ars).

model (R > 0.99). DSC scans of PI3-SH3 exhibited a single cooperative transition corresponding to a two-state unfolding process without a significant population of intermediate partially unfolded states, according to the ratio between the van't Hoff and calorimetric enthalpies (Fig. 4C). As shown in Table 1, the transition temperatures of unfolding obtained by the three independent methods are very similar, coinciding to indicate that PI3-SH3-SS is considerably more stable toward thermal denaturation than the wild-type and reduced forms. PI3-SH3-$(SH)_2$ is somewhat more stable than the wild-type domain.

The temperature-induced unfolding of PI3-SH3-WT and PI3-SH3-SS was also analyzed by monitoring the changes in their ^{1}H-NMR spectra (Fig. 3). PI3-SH3-WT displays native signal dispersion and peak sharpness until 318 K; above this temperature the intensity of the signals steadily decays until at 338 K the spectrum collapses and the resonances at low fields are hardly detectable, indicating the absence of a preferential folded conformation. In the case of PI3-SH3-SS, the spectra remain essentially unchanged until 328 K and even at 338 K the spectrum presents good signal dispersion, although with decreased peak intensity. Both proteins appear to be completely unfolded at 358 K (data not shown). The area of the NMR signals in the aliphatic region was plotted against temperature and the resultant curves were adjusted to a two-state model (Fig. 4D). The derived melting temperatures confirm the stabilization of the PI3-SH3-SS domain against thermal unfolding (Table 1).

Equilibrium denaturation of PI3-SH3 domains

The urea denaturation at equilibrium of the different PI3-SH3 domains was analyzed at pH 7.0 and 298 K. Protein unfolding was monitored by following the changes in intrinsic fluorescence at increasing urea concentrations. All three domains display a single detectable transition indicating the cooperativity of the unfolding process (Fig. 5A). The main thermodynamic parameters of the unfolding reaction were calculated from the equilibrium curves assuming a two-state model ($R > 0.99$ in both cases) (Table 2). The stability of PI3-SH3-WT is 4.4 kcal/mol with a $[\text{urea}]_{50\%}$ of 3.8 M. In agreement with the thermal denaturation data, PI3-SH3-$(SH)_2$ is slightly more stable. Both domains display similar m-values. The value of $[\text{urea}]_{50\%}$ for PI3-SH3-SS is ~7.5 M; this value is so high that a final baseline for the fully unfolded state cannot be accurately measured and the m-value could not be determined. Therefore, we used guanidine hydrochloride (Gdn·HCl) for equilibrium denaturation of PI3-SH3-SS and compared it with that of the two other domains using the same denaturant (Fig. 5B). As detailed in Table 2, crosslinking stabilizes PI3-SH3-SS by ~3.5 kcal/mol relative to PI3-SH3-WT and shifts $[\text{Gdn·HCl}]_{50\%}$ from 1.7 to

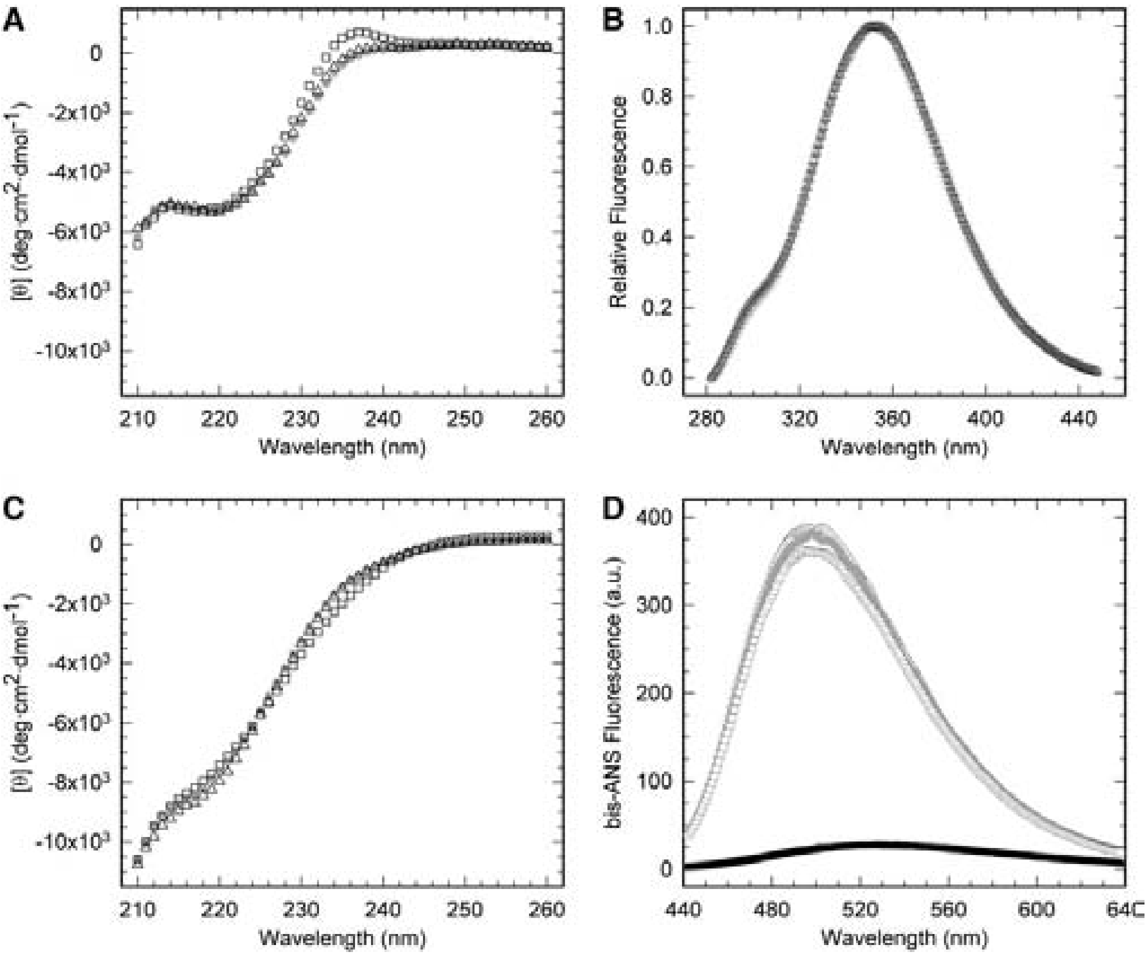


FIG. 2. Conformational analysis of PI3-SH3 domains. (A) Far-UV circular dichroism (CD) spectra at pH 7. **(B)** Intrinsic fluorescence emission spectra upon excitation at 268 nm at pH 7. **(C)** Far-UV CD spectra at pH 2. **(D)** 4,4′-dianilino-1,1′-binaphthyl-5,5′-disulfonic acid (bis-ANS) binding at pH 2. PI3-SH3-WT (*squares*), PI3-SH3-SS (*solid gray circles*), and PI3-SH3-$(SH)_2$ (*triangles*). In **(D)** bis-ANS binding to SH3 domains under native conditions is indicated with *solid black symbols*.

2.9 *M*. Overall, thermal and chemical denaturation data coincide to indicate that the introduced 3–82 disulfide bond is strongly stabilizing both relative to the wild-type and the dithiol forms and therefore that the mutations are sterically acceptable.

Folding and unfolding kinetics of PI3-SH3 domains

The kinetics of folding and unfolding of the different PI3-SH3 domains were determined by stopped-flow at neutral pH and 298 K under a wide range of denaturant conditions. In all cases, the folding and unfolding traces by fluorescence fit well into single exponentials, indicating the lack of detectable intermediates according with a two-state model. The chevron plots for the different domains appear to be linear in the complete range of urea concentrations studied (Fig. 6). The rate constants for folding (k_f) and unfolding (k_u) and their dependence on the denaturant concentration (m_f and m_u) are shown in Table 3. The kinetic data agree with the equilibrium stabilities. Interestingly, PI3-SH3-SS folds over 35 times faster than PI3-SH3-WT (Fig. 6). This high acceleration of the folding reaction can be univocally attributed to the presence of the covalent bond, since PI3-SH3-$(SH)_2$ exhibits a k_f close to that of the wild-type domain. In contrast, cyclization has only a minor effect on the unfolding rate. The calculated $\Phi_{F\text{-}U}$ value of 0.85 for the disulfide crosslink suggested that the N and C termini might be ordered in the transition state for folding. We analyzed the folding and unfolding kinetics of Y8A and V74A mutants in the N- and C-terminal β-strands to probe the degree of association between PI3-SH3-WT ends in the transition state (Fig. 6B). The side-chains of Tyr8 and Val47 face each other in the native structure. The two mutations are highly destabilizing (Table 3), indicating that these residues play an important role in maintaining the native conformation. However, the $\Phi_{F\text{-}U}$ values of 0.16 and 0.20 calculated for Y8A and V74A, respectively, indicate that, in the wild-type domain, the terminal β-strands 1 and 5 are not part of the folding nucleus.

Reductive unfolding and oxidative folding

In the presence of 1 m*M* dithiothreitol (DTT), the disulfide bond in native PI3-SH3-SS is completely reduced in less than 1 min, indicating that it is highly exposed to solvent (Supplementary Fig. S1A; Supplementary Data are available online at www.liebertonline.com/ars). *In vitro*, the oxidative folding reaction of PI3-SH3-$(SH)_2$ from denatured state is very inefficient,

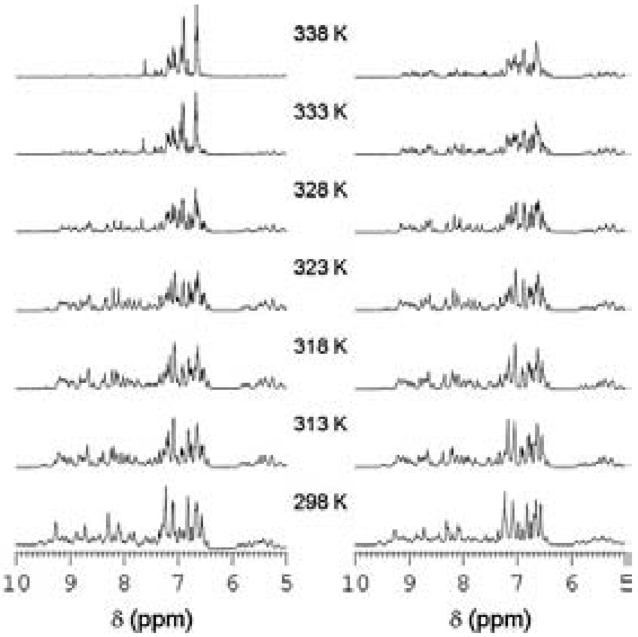


FIG. 3. ^{1}H-NMR spectra of PI3-SH3-WT (left) and PI3-SH3-SS (right) at the indicated temperatures.

taking more than 120 min even in the presence of an oxidizing agent such as 1 m*M* GSSG (Supplementary Fig. S1B). As shown above, the conformational folding of PI3-SH3-(SH)$_2$ occurs in seconds, indicating that the formation of the disulfide bond is the rate-limiting step of the oxidative folding reaction. This suggests that, in the reduced state, the free cysteine residues possess high degree of flexibility, which renders more difficult their oxidation to form the intramolecular disulfide bond.

Conformational analysis of PI3-SH3 domains at low pH

PI3-SH3-WT readily forms amyloid-like fibrils *in vitro* when incubated under acidic conditions. At pH 2.0 the CD spectra of all three domains display transitions to more unfolded conformations (Fig. 2C). These transitions do not reflect global unfolding, as when adding a strong chemical denaturant, but rather partial unfolding, since it is known that at pH 2.0 PI3-SH3-WT forms the so-called A-state, a conformation that is more compact that the denatured state but less than the native one (28, 46). The A-state contains little secondary structure and exposes hydrophobic surfaces to solvent; a property that can be monitored using 4,4′-dianilino-1,1′-binaphthyl-5,5′-disulfonic acid (bis-ANS). The A-state, from which amyloid assembly starts, seems to be accessible to the three domains, since in all cases we could detect a significant increase in the fluorescence emission of bis-ANS and a blue shift of its λ_{max} in the presence of the proteins at low pH (Fig. 2D). The native domains had a negligible effect on the spectral properties of the dye.

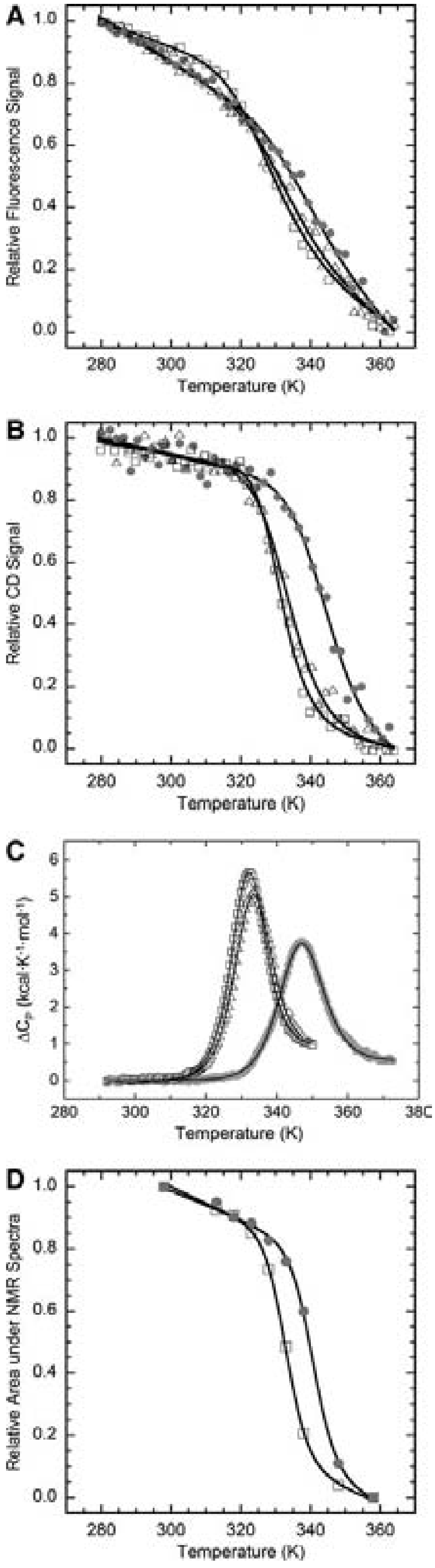


FIG. 4. Thermal unfolding followed by (A) intrinsic fluorescence, (B) far-UV CD, (C) differential scanning calorimetry, and (D) ^{1}H-NMR of PI3-SH3-WT (*squares*), PI3-SH3-SS (*solid gray circles*), and PI3-SH3-(SH)$_2$ (*triangles*). Nonlinear fit curves are represented by continuous lines.

Table 1. Thermal Denaturation Thermodynamic Parameters

	ΔH (T_m) *(kcal/mol)*[a]	*ΔC*$_p$ *(kcal/K·mol)*[a]	T_m *(K)*[a]	T_m *(K)*[b]	T_m *(K)*[c]	T_m *(K)*[d]
PI3-SH3-WT	67 ± 3	0.9 ± 0.2	331.2 ± 0.5	331.1 ± 0.3	331.9 ± 0.1	332.2 ± 0.2
PI3-SH3-SS	57 ± 3	0.5 ± 0.2	346.0 ± 0.5	345.2 ± 2.1	345.1 ± 0.2	342.1 ± 0.6
PI3-SH3-$(SH)_2$	63 ± 3	0.9 ± 0.2	332.4 ± 0.5	334.9 ± 0.4	333.9 ± 0.1	—

[a] DSC.

[b] Intrinsic fluorescence.

[c] CD.

[d] 1H-NMR.

ΔC_p, heat capacity change of unfolding; CD, circular dichroism; DSC, differential scanning calorimetry; PI3-SH3, SH3 domain of p85α subunit of bovine phosphatidyl-inositol-3′-kinase.

Sequential aggregation propensity of PI3-SH3 domains

We used AGGRESCAN (18) and WALTZ (35) algorithms to estimate the effect of mutations on the intrinsic aggregation propensity of PI3-SH3. AGGRESCAN predicts three identical aggregation-prone regions for the wild-type and mutant proteins, comprising residues G27-V32, G35-F42, and G71-I77 (Fig. 1A). The first and last regions correspond with the second and last β-strands and the second region to the α-helix at the N-src loop. This algorithm calculates aggregation propensities of −19.9 and −20.7 for PI3-SH3-WT and PI3-SH3-$(SH)_2$, respectively. WALTZ identifies four identical amyloid-prone regions in both domains (Fig. 1A), comprising residues G5-Y12 in β-strand 1, G27-G41 in the RT-loop, E52-N60 in the N-src loop, and β-strand 3 and F69-G78 in the last β-strand. Overall, the introduced mutations are not expected to promote by themselves large changes in the aggregation propensity of the domain.

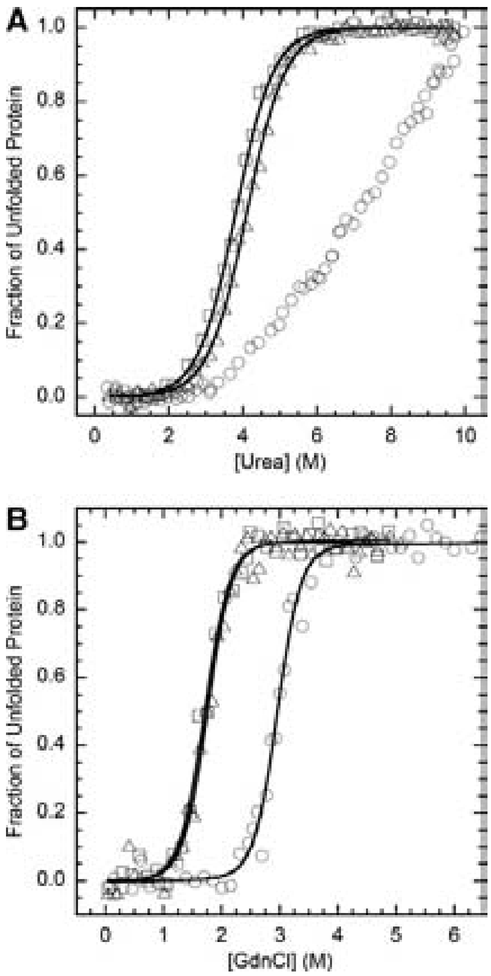


FIG. 5. Equilibrium unfolding in the presence of (A) urea and (B) guanidine hydrochloride (Gdn·HCl) of PI3-SH3-WT (*squares*), PI3-SH3-SS (*circles*), and PI3-SH3-$(SH)_2$ (*triangles*).

Amyloid fibril formation by PI3-SH3 domains

We analyzed the kinetics of PI3-SH3 domains amyloid fibril formation at acidic pH by monitoring the changes over time in the fluorescence of the amyloid staining dye Th-T (Fig. 7). In all cases we obtained characteristic polymerization sigmoidal curves, reflecting a nucleation-polymerization process that can be analyzed as an autocatalytic reaction (40). In agreement with the theoretical predictions, PI3-SH3-WT and PI3-SH3-$(SH)_2$ aggregation curves overlap exhibiting similar nucleation ($7.10 \times 10^3\ s^{-1}$ and $6.29 \times 10^3\ s^{-1}$, respectively) and elongation rates ($1.76 \times 10^{-3}\ M^{-1}\ s^{-1}$ and $1.02 \times 10^{-3}\ M^{-1}\ s^{-1}$, respectively). In contrast, the aggregation reaction of PI3-SH3-SS is significantly slower, with a 7-fold slower nucleation rate ($1.01 \times 10^3\ M^{-1}\ s^{-1}$) and 4-fold slower elongation rate ($4.09 \times 10^{-4}\ M^{-1}\ s^{-1}$) than the wild-type protein.

Analysis of the different aggregation reactions by transmission electron microscopy (TEM) (Fig. 8) shows that the formation of protofibrilar aggregates in PI3-SH3-SS samples is

Table 2. Thermodynamic Properties of Equilibrium Unfolding of PI3-SH3 Domains

Domain/denaturant	*[]$_{50\%}$ (M)*	*ΔG$_{U\text{-}F}$ (kcal/mol)*	*m (kcal/molM)*	*ΔΔG$_{U\text{-}F}$*
PI3-SH3-WT/Urea	3.8	4.43 ± 0.22	1.16 ± 0.05	—
PI3-SH3-$(SH)_2$/Urea	4.1	4.67 ± 0.20	1.14 ± 0.05	0.24
PI3-SH3-SS/Urea	~7.5	—	—	—
PI3-SH3-WT/GdmCl	1.7	4.76 ± 0.81	2.74 ± 0.42	—
PI3-SH3-$(SH)_2$/GdmCl	1.8	4.92 ± 0.71	2.80 ± 0.36	0.16
PI3-SH3-SS/GdmCl	2.9	8.34 ± 1.13	2.81 ± 0.37	3.58

$\Delta G_{U\text{-}F}$ (free energy of unfolding) and $\Delta\Delta G_{U\text{-}F}$ (the difference in $\Delta G_{U\text{-}F}$ between the particular mutant protein and WT) were calculated from the equilibrium parameters as described in the Materials and Methods section.

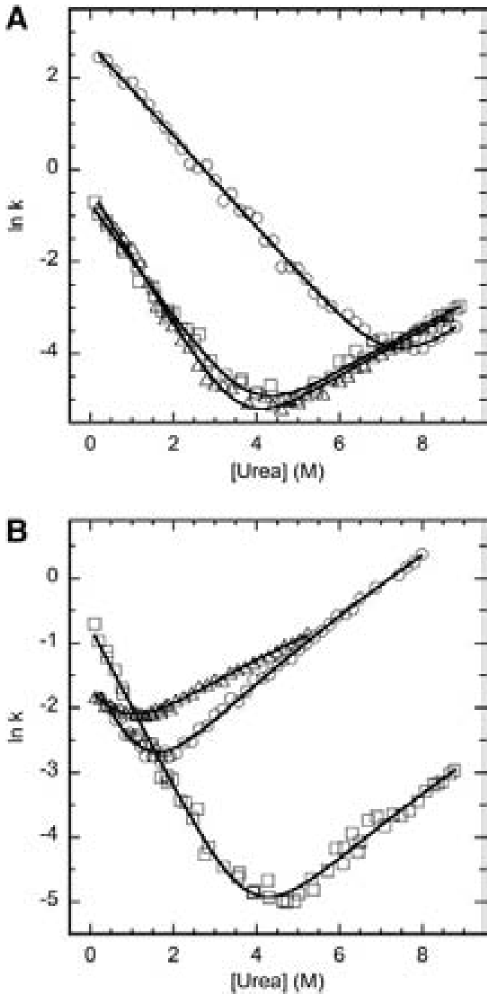


FIG. 6. Folding and unfolding kinetics. (A) Dependence of the folding and unfolding rate constants (*k*) on urea concentration of PI3-SH3-WT (*squares*), PI3-SH3-SS (*circles*), and PI3-SH3-$(SH)_2$ (*triangles*). **(B)** Dependence of the folding and unfolding rate constants (*k*) on urea concentration of PI3-SH3-WT (*squares*), Y8A (*triangles*), and V74A (*circles*) mutants.

delayed relative to that in PI3-SH3-WT and PI3-SH3-$(SH)_2$ solutions. After 42 h incubation, the three domains form typical amyloid fibrils (Fig. 8). However, differences in the morphology of the fibrils were observed. PI3-SH3-WT and PI3-SH3-$(SH)_2$ fibrils were long and eventually tended to twist around each other, whereas PI3-SH3-SS formed shorter, discrete, and straight structures (Fig. 8). No evident depolymerization or shift in morphology could be observed when mature PI3-SH3-SS fibrils were incubated for 5 h under reducing conditions (data not shown).

Conformational and toxicity properties of PI3-SH3 amyloid fibrils

The conformational stability of the fibrils formed by the different domains after 42 h was assessed by equilibrium chemical denaturation with Gdn·HCl at pH 2.0 and 298 K following the changes in Th-T fluorescence (Fig. 9A) and light scattering (Fig. 9B). The fibrils seem to denature in a cooperative manner. The $[Gdn \cdot HCl]_{50\%}$ calculated by measuring Th-T fluorescence were 3.2, 2.4, and 2.2 for PI3-SH3-SS, PI3-SH3-$(SH)_2$, and PI3-SH3-WT fibrils, respectively. Light scattering measurements rendered $[Gdn \cdot HCl]_{50\%}$ values of 3.9, 3.1, and 2.9 for PI3-SH3-SS, PI3-SH3-$(SH)_2$, and PI3-SH3-WT fibrils, respectively. Therefore, the fibrils of PI3-SH3-SS appear to be more stable than those of the wild-type and reduced forms. In all cases, the stability calculated from Th-T measurements is lower than that obtained from light scattering. The difference in the two transition midpoints suggests the presence of intermediate states in which the ordered β-sheet is disrupted, in such a way that it does not longer bind Th-T, but still remains in an aggregated state, contributing thus to the light scattering signal. The population of these species can be approximated from the normalized denaturation data and fitted to a Gaussian distribution (Fig. 9C). The intermediate species distributions suggest that the β-sheet structure in the wild-type fibrils is less stable than in the fibrils of the disulfide crosslinked domain. To further explore whether the observed differences in fibrils morphology and stability are related to differences in the interactions supporting their architecture, we analyzed the conformational properties of PI3-SH3 mature fibrils by attenuated total reflectance–Fourier transformed infrared spectroscopy (ATR-FTIR) and bis-ANS binding and limited proteolysis. The ATR-FTIR spectra of all three fibrils in the amide I region is dominated by a main band at 1615–1640 cm^{-1} corresponding to β-sheet conformations (Fig. 9D–F). The secondary structure composition of PI3-SH3-WT (Fig. 9D) and PI3-SH3-$(SH)_2$ (Fig. 9E) fibrils are almost identical and differ from that of PI3-SH3-SS fibrils (Fig. 9F). The proportion of amyloid-like intermolecular β-sheet secondary structure is higher in the fibrils formed by PI3-SH3-SS (72%) than in those

Table 3. Folding Kinetic Parameters for PI3-SH3 Domains

Domain	$k_f^{0.5M}$ *(s^{-1})*	k_u^{8M} *(s^{-1})*	m_f *(kcal/molM)*	m_u *(kcal/molM)*	$-RT_{mF\text{-}U}$ *(kcal/molM)*	$\Delta G_{U\text{-}F}$ *(kcal/mol)*	$\Delta\Delta G_{U\text{-}F}$
PI3-SH3-WT	0.256 ± 0.013	0.0825 ± 0.0102	1.28 ± 0.03	0.51 ± 0.02	1.07 ± 0.03	4.39	—
PI3-SH3-$(SH)_2$	0.311 ± 0.019	0.0794 ± 0.0099	1.37 ± 0.03	0.58 ± 0.02	1.15 ± 0.03	4.62	0.23
PI3-SH3-SS	9.312 ± 0.389	0.0472 ± 0.0356	1.03 ± 0.03	0.75 ± 0.10	1.05 ± 0.07	8.04	3.65
Y8A	0.135 ± 0.005	0.8537 ± 0.0286	1.35 ± 0.11	0.45 ± 0.08	1.07 ± 0.10	0.09	−4.30
V74A	0.128 ± 0.007	1.4331 ± 0.0596	1.44 ± 0.08	0.67 ± 0.08	1.25 ± 0.08	1.52	−2.87

Kinetics of folding and unfolding were followed by changes in intrinsic fluorescence on a stopped flow instrument at 298 K; k_f is reported in 0.5 *M* urea and k_u is in 8 *M* urea to avoid extrapolation; m_f and m_u are the dependences of the folding and the unfolding rates, respectively, on urea. $\Delta G_{U\text{-}F}$ (free energy of unfolding) and $\Delta\Delta G_{U\text{-}F}$ (the difference in $\Delta G_{U\text{-}F}$ between the particular mutant protein and WT) were calculated from the kinetic parameters as described in the Materials and Methods section.

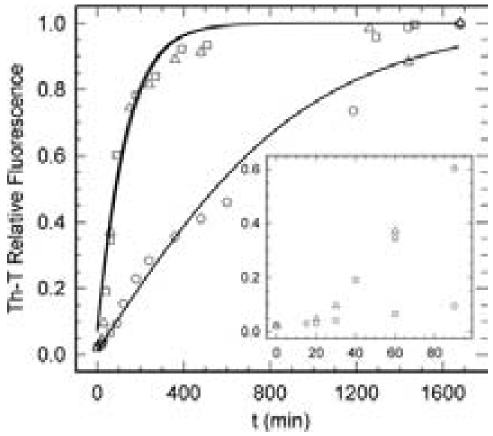


FIG. 7. Aggregation kinetics under acidic conditions of PI3-SH3-WT (*squares*), PI3-SH3-SS (*circles*), and PI3-SH3-(SH)$_2$ (*triangles*). The *inset* corresponding to the first 100 min shows the aggregation lag phase for the three domains.

formed by PI3-SH3-WT (61%) and PI3-SH3-(SH)$_2$ (60%) (Supplementary Table S1). In agreement with these data, PI3-SH3-WT and PI3-SH3-(SH)$_2$ fibrils exhibit higher binding to bis-ANS than the fibrils formed by the disulfide-containing domain, indicating the existence of a lower proportion of hydrophobic residues exposed to solvent in the fibrils formed by PI3-SH3-SS (Fig. 9G). Accordingly, we observed that, in our conditions, the PI3-SH3-SS fibrils were totally resistant to pepsin digestion, whereas the wild-type fibrils were susceptible to proteolysis (Fig. 9H).

The toxicity of amyloid fibrils is related to their conformational properties. It has been shown for different and unrelated proteins that the binding to ANS-like dyes correlates with the toxicity of amyloid species, suggesting that the exposure of hydrophobic regions is a critical characteristic of these pathogenic assemblies (5). Although the aggregation of PI3-SH3 is not associated to any known disease, its amyloid assemblies are inherently cytotoxic (7). We analyzed the toxicity of PI3-SH3 fibrils on cultured neuroblastoma cells from the SH-SY5Y cell line. In excellent agreement with bis-ANS binding data, the fibrils formed by PI3-SH3-SS were less toxic than those formed by the PI3-SH3-WT and PI3-SH3-(SH)$_2$ domains (Fig. 9I).

Extracellular SH3 domains display disulfide bonds and a high sequential intrinsic aggregation propensity

SH3 domains have been considered traditionally as structural elements of multidomain intracellular signaling proteins and devoid of disulfide bonds. However, it has been shown recently that the extracellular melanoma inhibitory activity protein (MIA) consists of a single domain adopting an SH3-fold covalently linked by two intramolecular disulfide bonds. Moreover, three MIA homologous extracellular

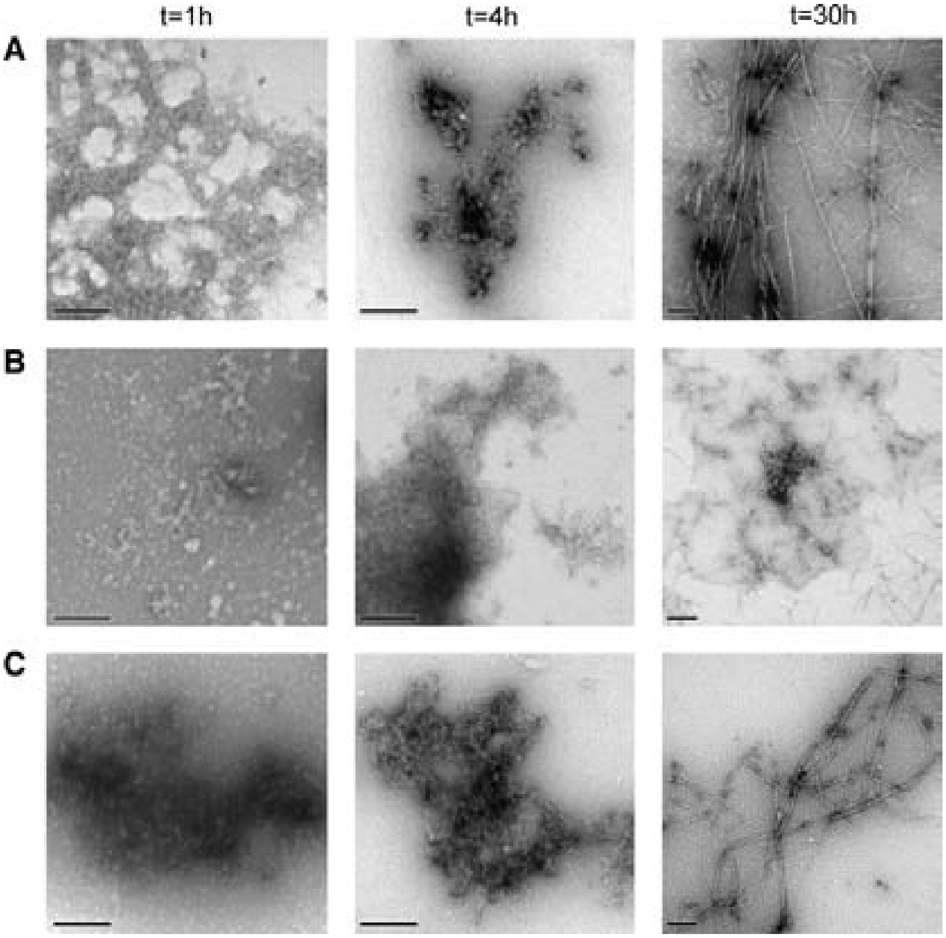


FIG. 8. Transmission electron micrographs of (A) PI3-SH3-WT, (B) PI3-SH3-SS, and (C) PI3-SH3-(SH)$_2$ negatively stained samples collected at different times along aggregation experiments. The scale bar represents 200 nm.

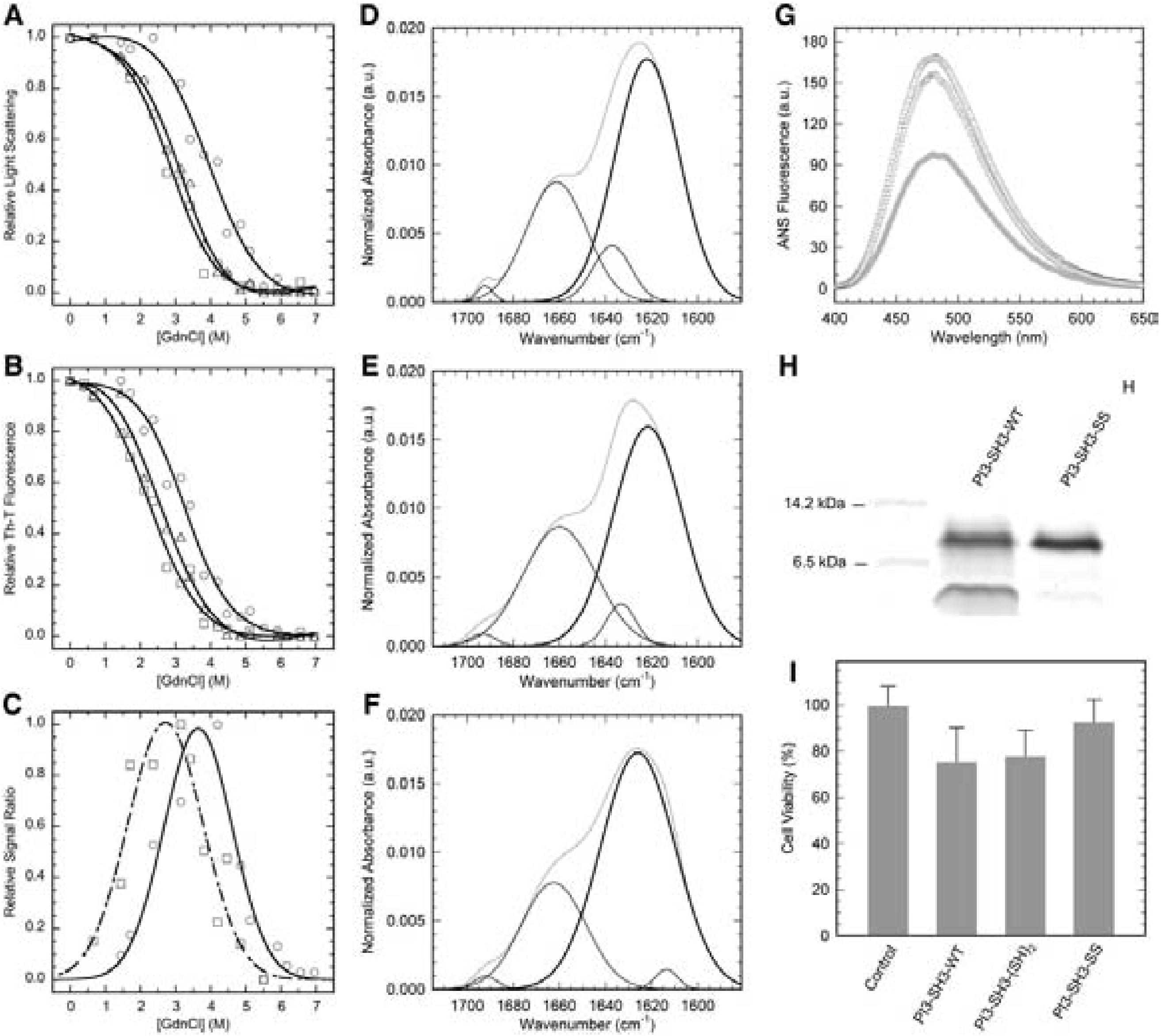


FIG. 9. Conformational and toxic properties of PI3-SH3 domains. (A) Fibrils stability in the presence of Gdn·HCl followed by light scattering. **(B)** Fibrils stability followed by thioflavin-T binding. **(C)** Population of intermediate states. **(D–F)** Attenuated total reflectance–Fourier transformed infrared spectroscopy spectra in the amide I region of PI3-SH3-WT, PI3-SH3-$(SH)_2$, and PI3-SH3-SS (respectively); the thick line indicates the contribution of intermolecular β-sheet to the spectra. **(G)** ANS binding to PI3-SH3 fibrils. **(H)** SDS-PAGE analysis of the proteolytic susceptibility of amyloid fibrils. **(I)** Viability analysis of SH-SY5Y cells exposed to PI3-SH3 amyloid fibrils for 48 h. Error bars indicate ± SE ($n = 4$). One hundred percent cell viability was assigned to control samples corresponding to cells incubated in domain free DMEM. In **(A, B, C,** and **G)** *panels* symbols correspond to PI3-SH3-WT (*squares*), PI3-SH3-SS (*circles*), and PI3-SH3-$(SH)_2$ (*triangles*).

proteins with the four conserved Cys have been identified (TANGO, MIA-2, and OTOR) (44). Since our data suggest that the presence of disulfide bonds might modulate the aggregation propensity of SH3 domains, we compared the intrinsic aggregation properties of these extracellular SH3 disulfide-containing domains with those of classical intracellular ones using AGGRESCAN (Fig. 10). The analysis indicates that extracellular SH3 domains display a higher number of aggregation-prone regions in their sequence and support a higher intrinsic aggregation load than intracellular domains.

Discussion

Cyclization increases PI3-SH3 domain stability

Biophysical data indicate that the three PI3-SH3 domains analyzed in the present study exhibit a properly folded conformation in their native states. The dithiol form is slightly more stable than the wild-type domain. DSC data indicate that ΔH is lower, in absolute value, in PI3-SH3-$(SH)_2$ than in PI3-SH3-WT and therefore that stabilization does not arise from the establishment of favorable interactions by the thiol groups but rather from a decrease in ΔS in the mutant form.

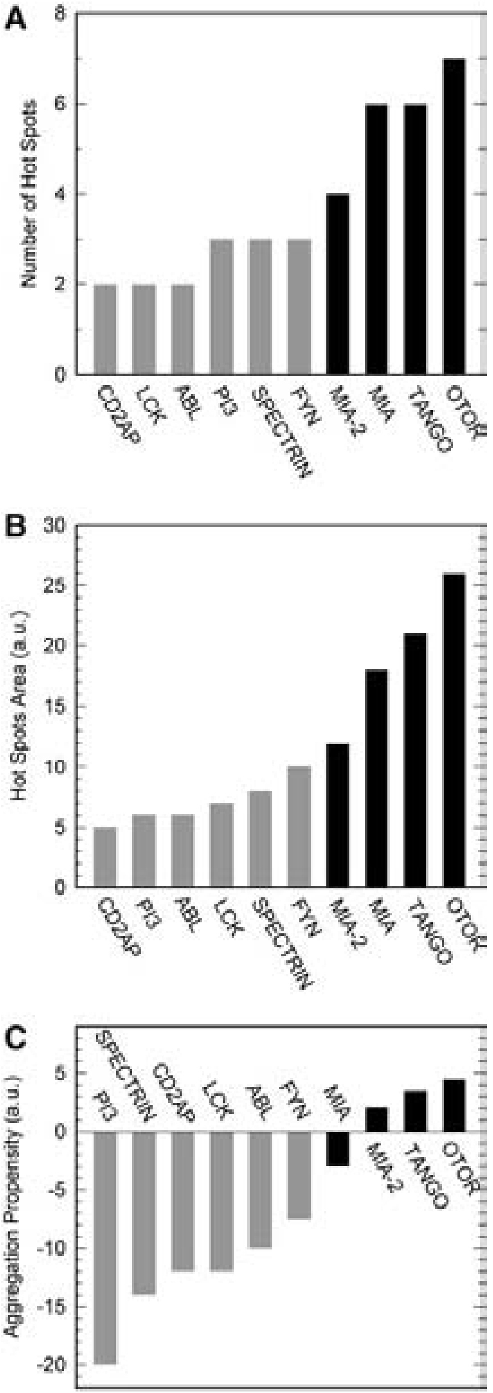


FIG. 10. Intrinsic aggregation propensities of intracellular (gray) and extracellular (black) SH3 domains. All extracellular domains share two conserved disulfide bonds. **(A)** Number of aggregation-prone regions in the sequences. **(B)** Area of aggregation-prone regions in the AGGRESCAN profile, reflecting the theoretical aggregation potency of these regions. **(C)** Predicted aggregation propensities of SH3 domains. Higher values indicate higher aggregation propensity.

The ProtSA server (21) indicates that, in the wild-type domain, Ala3 and Ile82 are more exposed to solvent than expected for these hydrophobic residues. Their mutation to Cys displaying polar free thiols in the reduced state would be entropically favorable and might well account for the observed increase in stability. Oxidative folding experiments confirm that the Cys are exposed to solvent and in a flexible environment in PI3-SH3-(SH)2.

Direct comparison of PI3-SH3-SS with its dithiol form allows differentiating the effect of the mutations from the effect of the disulfide bond. Crosslinking of PI3-SH3 stabilizes the domain against both thermal and chemical denaturation, with an increase in T_m of $\sim$15 K and a $\Delta\Delta G_{U\text{-}F}$ at 298 K of $\sim$3.5 kcal/mol, relative to the dithiol form. It is generally assumed that the stability provided by a disulfide bond results largely from a reduction in the configurational entropy (ΔS_{conf}) of the unfolded state. The ΔS_{conf} is related to the size of the polypeptide loop enclosed by the disulfide bond (N) and can be approximated theoretically according to the equation (37):

$$\Delta S_{conf} = 2.1\ (3/2)\ \mathrm{R} \ln \mathrm{N} \qquad (1)$$

where R is the gas constant.

According to Equation (1) the expected stabilization of PI3-SH3-SS, relative to its reduced form, is 4.5 kcal/mol. Therefore, although the experimentally measured increase in stability is among the highest reported for an SH3 domain (27) it is lower than the theoretical estimate, indicating that loop entropy is not the only factor determining the stability of the cycled domain. Together with a decrease in ΔS, as measured by DSC, the disulfide bond causes a decrease in ΔH, indicating enthalpy–entropy compensation. The same effect has been observed for a number of disulfide mutants, including those of a camelid antibody (29), cytochrome c (4), and barnase (31), demonstrating that the precise thermodynamic effects of introducing a covalent link in a polypeptide are hardly predictable from simple theoretical considerations (16). In these protein models, disulfide bonds tend to decrease the heat capacity change of unfolding (ΔC_p). The ΔC_p between the folded and unfolded states of proteins is thought to arise from the exposure of buried side-chains to solvent, with minor contributions from changes in configurational entropy. As proposed by Doig and Williams (20), the 45% decrease in the ΔC_p of PI3-SH3-SS likely results from a more compact unfolded state with restricted hydration of side chains.

Cyclization accelerates PI3-SH3-folding

Cyclization is always expected to increase folding rates due to the larger decrease in configurational entropy of the unfolded state relative to that in the transition state. Unfolding rates will decrease only if the polypeptide termini become apart at the transition state. Connection of the N- and C- termini in PI3-SH3-SS causes a large increase, by over 1 order of magnitude, of the folding rate and has, by contrast, only a small influence on the unfolding rate. This behavior has also been observed for disulfide mutants of barnase (15), subtilisin (45), and acylphosphatase (38). However, cyclization in the structurally homologous SH3 domain of src kinase (src-SH3) domain affected roughly equally its folding and unfolding rates (27), suggesting that its ends are not as structured in the transition state as they are in the native conformation and, likely, that a rate-limiting step in unfolding is the dissociation of the N- and C- termini (27). This view is consistent with the results obtained by Φ analysis for this protein (39) and the spectrin-SH3 domain (34). The two studies indicate that the folding transition state of these domains involves the association of the distal β-hairpin and the diverging turn, whereas the N and C termini are completely disordered. The low Φ values obtained here for mutants in the β-strands 1 and 5 indicate that the protein ends are also unstructured in the transition state of PI3-SH3-WT, confirming thus a robust energy landscape for the folding of SH3 domains. Cyclization makes the topology of the protein symmetric and tends to reduce the polarization of the transition state. The dissimilar Φ values obtained by circularization and point mutation of N- and C- regions in PI3-SH3 indicate that the free energy landscape of this particular domain is severely affected by the disulfide bond promoted topology change, confirming that despite the folding landscape of this protein family is relatively insensitive to sequential variations it is sensitive to topological constrains (26).

Cyclization slows down PI3-SH3 aggregation

The effect of polypeptide chain crosslinking in amyloid fibril formation is still poorly understood. Disulfide crosslinking does not prevent self-association into amyloid structures. Therefore, the conformational restrictions introduced by the covalent bond do not prevent the accommodation of the polypeptide backbone into the highly ordered cross-β-sheet structure characteristic of amyloid fibrils. However, it strongly decreases the nucleation and elongation rates, slowing down the lag phase as well as the overall aggregation reaction. The aggregation of PI3-SH3-WT has been shown to be consistent with the nucleated conformational conversion mechanism (8), in which initially soluble monomers coalesce to form amorphous assemblies that later reorganize into ordered oligomers and finally β-sheet enriched amyloid fibrils (43). Hydrophobic forces are thought to drive the initial unspecific condensation of monomers into amorphous oligomers. According to the conformational properties and hydrophobic residues clustering of PI3-SH3-SS and PI3-SH3-(SH)$_2$ domains, there is not obvious reason to assume differences in their association rates at this stage. Upon the initial collapse, polypeptides within the disordered oligomers realign slowly to establish hydrogen bonds that favor the formation of β-sheets. The dynamic flexibility of the polypeptide chain might play an important role in this second step, since multiple rounds of association and dissociation between adjacent protein regions likely occur during the lag phase, before an initially ordered and stable enough β-sheet conformation is attained (2). The reduced flexibility of PI3-SH3-SS might difficult this rearrangement stage delaying the formation of productive oligomers, resulting in a lower k_n.

The elongation step in amyloid formation is thought to involve a "dock and lock" mechanism, in which a significant conformational rearrangement of incoming monomers is necessary before they become incorporated into preformed fibrils (17). The slower elongation rate of PI3-SH3-SS suggests that dynamic protein flexibility might be also important at the polymerization stage.

Cyclization reduces the toxicity of PI3-SH3 amyloid assemblies

High-resolution MAS NMR analysis has allowed Bayro *et al*. to propose a structural model for PI3-SH3-WT amyloid fibrils (3). In the model, four β-sheets form the core of two distinct protofilaments and lateral contacts between PI3-SH3-WT subunits in adjacent protofilaments allow their association to form the fibrils (Supplementary Fig. S2). The N-terminus adopts a rigid β-strand structure in the core of the fibril. In contrast, the C-terminal residues appear to be dynamically and structurally disordered and located in the two external sides of the fibril (3). Restriction of the C-end entropy by crosslinking might allow its integration in the core β-sheet structure (Supplementary Fig. S2). This would explain the highest stability of the β-sheet structure in PI3-SH3-SS fibrils, the insensitivity of these fibrils to highly reducing conditions, the higher proportion of intermolecular β-sheet and the resistance of fibrils to proteolysis, as well as the reduced exposition of hydrophobic residues to solvent. This cyclization-promoted structural gain has a crucial functional consequence: it significantly reduces the cytotoxicity of the resulting fibrils.

Materials and Methods

Disulfide bond design

A PI3-SH3 domain with the N and C terminal regions linked by a disulfide bridge (PI3-SH3-SS) was designed using FoldX (version 2.65) (http://foldx.crg.es/) (42). The best positions to introduce the Cys residues were identified employing the BuildModel command, which calculates the relative stability difference by rotating the same residues in the wild-type and mutant structures according to the FoldX rotamer database. The atomic spatial coordinates of PI3-SH3-WT in the crystal structure (1PHT.pdb) were used as input.

Cloning, mutagenesis, and expression

The PI3-SH3-WT protein consists of 83 residues from the SH3 domain of bovine PI3-kinase plus a Met at the N-terminus. Its cDNA was cloned in pBAT-4 (46). The Ala3 and Ile82 residues were mutated to Cys, the resulting plasmid was transformed in BL21(DE3) *Escherichia coli* cells and the proteins were expressed and purified as described previously (46). The mutant cDNA was sequenced and the protein identity was checked by mass spectrometry. Gel filtration chromatography indicated that the mutant protein is a monomer. Unless otherwise indicated, experiments were performed in 50 m*M* sodium phosphate at pH 7.0 (buffer-N).

Disulfide crosslinking and reduction

To confirm the formation of the designed disulfide bond, PI3-SH3-SS was incubated in 0.1 *M* Tris-HCl buffer, pH 8.5, containing 0.1 *M* 4-vinylpyridine for 1 h at room temperature, diluted 1:10 in 0.1% aqueous trifluoroacetic acid, and analyzed by MALDI-TOF MS in an Ultraflex spectrometer (Bruker). When required, the disulfide bond was reduced by incubation in the presence of 10 m*M* DTT or Tris (2-carboxyethyl)phosphine (TCEP) for at least 1 h.

Oxidative folding and reductive unfolding

For oxidative folding experiments PI3-SH3-SS was reduced and unfolded in 0.5 *M* Tris buffer pH 8.4, containing 10 m*M* DTT and 7 *M* guanidinium chloride for 2 h at room temperature. To initiate folding, the sample was loaded onto a desalting column equilibrated with 0.1 *M* Tris buffer, pH 8.4. The protein was eluted in the same buffer to a final protein concentration of 0.3 mg/ml, in the absence and in the presence of 0.5 m*M* GSSG. Folding species were trapped in a time-course manner by alkylation with 0.1 *M* sodium iodoacetate for 2 min at room temperature. The trapped intermediates were subsequently analyzed by SDS-PAGE under non-denaturing conditions. For reductive unfolding experiments PI3-SH3-SS (0.5 mg/ml) was dissolved at room temperature in 0.1 *M* Tris buffer, pH 8.4, containing 1 m*M* DTT and the species were trapped and analyzed as described above.

Circular dichroism

CD spectra were measured in a Jasco-710 spectropolarimeter thermostated at 298 K. Spectra were recorded from 260 to 205 nm, at 1 nm intervals, 1 nm bandwidth, and a scan speed of 10 nm/min. Twenty accumulations were averaged for each spectrum. Protein concentrations were 20 μM in buffer-N. Thermal denaturation was monitored at 235 nm each 0.1 K, with 2 min temperature equilibrium between measures. Experimental data were fitted to a two-state transition curve for which the signals of the folded and unfolded states are dependent on the temperature using the nonlinear least squares algorithm provided with Kaleidagraph (Abelbeck Software).

Intrinsic fluorescence

Fluorescence was measured in a Varian Cary Eclipse spectrofluorometer using an excitation wavelength of 268 nm. Slit widths were typically 5 nm for excitation and 10 nm for emission. Spectra were acquired at 1 nm intervals, a 600 nm/min rate, and 0.1 s averaging time.

In PI3-SH3 domains, Trp55 is exposed to solvent and its fluorescence exhibits a linear dependence on the temperature or chaotropic agents concentration. Tyr residues are responsive to conformational changes, but their signal is strongly influenced by the fluorescence of Trp55. The ratio 303/350 nm provides an accurate measurement of structural changes in PI3-SH3 domains. Thermal denaturation was analyzed in buffer-N at 20 μM protein concentration each 0.1 K with 2 min equilibration between measures. For chemical denaturation the native protein was dissolved at 20 μM in buffer-N containing selected concentrations of denaturants (0–9 *M* urea or Gdn·HCl). The reaction was equilibrated for 20 h at room temperature. Experimental data were fitted as described for CD measurements.

Differential scanning calorimetry

The heat capacity of PI3-SH3 as a function of temperature was measured with a high-sensitivity differential scanning VP-DSC microcalorimeter (MicroCal). Protein samples and reference solutions were properly degassed and carefully loaded into the cells to avoid bubble formation. Thermal denaturation scans were performed with freshly prepared buffer-exchanged protein solutions. The baseline of the instrument was routinely recorded before the experiments. Experiments were performed in buffer-N, at a scanning rate of 1 K/min. Experiments were carried out with 40–50 μM of protein. Thermal denaturations were reversible (as judged by the agreement of the thermogram shapes and sizes between the first and the second scans). Pre- and post-transition baselines were considered linear with temperature.

A model-free analysis indicated that the calorimetric enthalpy is equal to the van't Hoff enthalpy, within experimental error. Therefore, protein stability parameters were obtained by analyzing thermal scans considering a two-state unfolding model using software developed in our laboratory implemented in Origin 7 (OriginLab). The unfolding thermodynamic parameters are not influenced by the ionization properties of the buffer as long as the pH of the experiment is close to the buffer pKa (7.2 for phosphate), and the buffer ionization enthalpy is small (0.86 kcal/mol for phosphate).

NMR spectroscopy

Samples were prepared at 40 μM in buffer-N, using a 9:1 H_2O/D_2O. One-dimensional NMR spectra were acquired at selected temperatures 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. The area under the NMR signals at the aliphatic region (form 10 to 5 ppm) was calculated and plotted to follow the temperature-induced denaturation of PI3-SH3 domains.

Determination of folding kinetic parameters

The kinetics of the folding and unfolding reactions at 298 K were followed in a Bio-Logic SFM-3 stopped-flow instrument using excitation at 268 and a 300 nm fluorescence cut-off filter. Starting with the protein in buffer-N, the unfolding reaction was promoted by dilution with appropriate volumes of the same buffer containing 9.5 *M* urea. For the folding reaction, appropriate volumes of urea-free buffer were added to an initial protein solution in buffer-N containing 9.5 *M* urea. Kinetic traces were fitted to single-exponential functions and the resulting rate constants to a two-state transition equation to determine the kinetic and free energy values:

$$\Delta G_{\text{F-U}} = \mathrm{R}T \ln(k_f/k_u)$$

where k_f and k_u are the rates of folding and unfolding, respectively, at denaturant concentrations experimentally accessible for the domain. The differences in the free energy of unfolding ($\Delta\Delta G_{\text{F-U}}$) between the wild-type protein and each mutant are calculated as:

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

$$\Delta\Delta G_{\ddagger\text{-U}} = \mathrm{R}T \ln(k_f^{wt}/k_f^{mut})$$

$$\Delta\Delta G_{\ddagger\text{-F}} = \mathrm{R}T \ln(k_u^{wt}/k_u^{mut})$$

The parameter Φ_F is defined as:

$$\Phi_F = \Delta\Delta G_{\ddagger\text{-U}}/\Delta\Delta G_{\text{F-U}}$$

and is interpreted as the fraction of the mutated residue's interactions that are formed in the transition state.

Amyloid fibril formation

Lyophilized PI3-SH3 domains were dissolved in buffer-N at 7 mg/ml. After filtration through a 0.22 μm filter to remove residual aggregates, protein solutions were diluted to 3.5 mg/ml adding an equal volume of 50 m*M* glycine, and the pH adjusted immediately with 5 *M* HCl to pH 2.0. Fibril formation was promoted by incubating the samples at 298 K and 500 rpm agitation. The 10 m*M* TCEP was used as reducing agent in these experiments.

Aggregation kinetics

After different incubation intervals samples were diluted in glycine buffer containing 60 μM Thioflavin-T (Th-T) and equilibrated 5 min at 298 K. They were excited at 450 nm and the fluorescence emission spectrum recorded between 470 and 600 nm with slit widths of 5 and 10 nm for excitation and emission, respectively. The kinetic parameters for the aggregation process were calculated according to an autocatalytic reaction, as previously described (40).

bis-ANS binding

The fluorescence emission of bis-ANS was recorded at 298 K. Protein samples (40 μM) at pH 7.0 or pH 2.0 were diluted tenfold into the corresponding buffer containing bis-ANS (2.6 μM). Spectra were collected after 5 min incubation. The samples were excited at 365 nm and emission measured between 440 and 640 nm with slit widths of 5 and 10 nm for excitation and emission, respectively.

Electron microscopy

Aggregated protein samples were diluted 20 times in the same buffer, 10 μl placed on a carbon-coated copper grid and allowed to stand for 5 min. The grid was washed with distilled water and the sample stained with 2% uranyl acetate for 1 min. The samples were imaged in a Hitachi H-7000 transmission electron microscope operating at an accelerating voltage of 75 kV.

Amyloid fibrils chemical denaturation

Mature fibril solutions were incubated in 0–7 *M* Gdn·HCl at pH 2 and for 16 h. Subsequently, they were diluted in glycine buffer containing 60 μM Th-T. The changes in light scattering at 340 nm and Th-T fluorescence at 480 nm were followed with a Varian Cary Eclipse spectrofluorimeter. Experimental data were fitted as described for CD measurements.

Limited proteolysis

Limited proteolysis was performed using pepsin in 50 m*M* glycine buffer at pH 2.0 and 298 K using an E/S ratio of 1:200 (by weight). The reactions were quenched after 5 min by adding an appropriate volume of 5 *M* NaOH. The proteolytic mixtures were analyzed by SDS-PAGE.

Attenuated total reflectance–Fourier transformed infrared spectroscopy

ATR-FTIR analyses of SH3 aggregates were performed using a Bruker Tensor 27 FTIR Spectrometer (Bruker Optics) with a Golden Gate MKII ATR accessory as previously described (41).

Cell viability assays

The toxicity of PI3-SH3 aggregates (10 μM) was analyzed on cultured neuroblastoma cells (SH-SY5Y cell line) by measuring formazan formation by mitochondrial dehydrogenases as previously described (41). Four independent assays were performed for each protein sample.

Acknowledgments

This work was supported by grants BFU2010-14901, BFU2010-19451, and BFU2010-16297 from Ministerio de Ciencia e Innovación (Spain), 2009-SGR-760 and 2009-CTP-00004 from AGAUR (Generalitat de Catalunya), and grupo consolidado *Protein Targets* (B89) (Gobierno de Aragón). S.V. has been granted an ICREA Academia award (ICREA).

Author Disclosure Statement

No competing financial interests exist.

References

1. Arolas JL, D'Silva L, Popowicz GM, Aviles FX, Holak TA, and Ventura S. NMR structural characterization and computational predictions of the major intermediate in oxidative folding of leech carboxypeptidase inhibitor. *Structure* 13: 1193–1202, 2005.
2. Auer S, Meersman F, Dobson CM, and Vendruscolo M. A generic mechanism of emergence of amyloid protofilaments from disordered oligomeric aggregates. *PLoS Comput Biol* 4: e1000222, 2008.
3. Bayro MJ, Maly T, Birkett NR, Macphee CE, Dobson CM, and Griffin RG. High-resolution MAS NMR analysis of PI3-SH3 amyloid fibrils: backbone conformation and implications for protofilament assembly and structure. *Biochemistry* 49: 7474–7484, 2010.
4. Betz SF and Pielak GJ. Introduction of a disulfide bond into cytochrome c stabilizes a compact denatured state. *Biochemistry* 31: 12337–12344, 1992.
5. Bolognesi B, Kumita JR, Barros TP, Esbjorner EK, Luheshi LM, Crowther DC, Wilson MR, Dobson CM, Favrin G, and Yerbury JJ. ANS binding reveals common features of cytotoxic amyloid species. *ACS Chem Biol* 5: 735–740, 2010.
6. Booth DR, Sunde M, Bellotti V, Robinson CV, Hutchinson WL, Fraser PE, Hawkins PN, Dobson CM, Radford SE, Blake CC, and Pepys MB. Instability, unfolding and aggregation of human lysozyme variants underlying amyloid fibrillogenesis. *Nature* 385: 787–793, 1997.
7. Bucciantini M, Giannoni E, Chiti F, Baroni F, Formigli L, Zurdo J, Taddei N, Ramponi G, Dobson CM, and Stefani M. Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature* 416: 507–511, 2002.
8. Carulla N, Zhou M, Arimon M, Gairi M, Giralt E, Robinson CV, and Dobson CM. Experimental characterization of disordered and ordered aggregates populated during the process of amyloid fibril formation. *Proc Natl Acad Sci U S A* 106: 7828–7833, 2009.
9. Cascales L and Craik DJ. Naturally occurring circular proteins: distribution, biosynthesis and evolution. *Org Biomol Chem* 8: 5035–5047, 2010.
10. Castillo V, Espargaro A, Gordo V, Vendrell J, and Ventura S. Deciphering the role of the thermodynamic and kinetic stabilities of SH3 domains on their aggregation inside bacteria. *Proteomics* 10: 4172–4185, 2010.
11. Castillo V, Grana-Montes R, Sabate R, and Ventura S. Prediction of the aggregation propensity of proteins from the primary sequence: aggregation properties of proteomes. *Biotechnol J* 6: 674–685, 2011.

12. Castillo V and Ventura S. Amyloidogenic regions and interaction surfaces overlap in globular proteins related to conformational diseases. *PLoS Comput Biol* 5: e1000476, 2009.
13. Chiti F and Dobson CM. Protein misfolding, functional amyloid, and human disease. *Annu Rev Biochem* 75: 333–366, 2006.
14. Chiti F and Dobson CM. Amyloid formation by globular proteins under native conditions. *Nat Chem Biol* 5: 15–22, 2009.
15. Clarke J and Fersht AR. Engineered disulfide bonds as probes of the folding pathway of barnase: increasing the stability of proteins against the rate of denaturation. *Biochemistry* 32: 4322–4329, 1993.
16. Clarke J, Hounslow AM, and Fersht AR. Disulfide mutants of barnase. II. Changes in structure and local stability identified by hydrogen exchange. *J Mol Biol* 253: 505–513, 1995.
17. Collins SR, Douglass A, Vale RD, and Weissman JS. Mechanism of prion propagation: amyloid growth occurs by monomer addition. *PLoS Biol* 2: e321, 2004.
18. Conchillo-Sole O, de Groot NS, Aviles FX, Vendrell J, Daura X, and Ventura S. AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics* 8: 65, 2007.
19. Creighton TE and Goldenberg DP. Kinetic role of a metastable native-like two-disulphide species in the folding transition of bovine pancreatic trypsin inhibitor. *J Mol Biol* 179: 497–526, 1984.
20. Doig AJ and Williams DH. Is the hydrophobic effect stabilizing or destabilizing in proteins? The contribution of disulphide bonds to protein stability. *J Mol Biol* 217: 389–398, 1991.
21. Estrada J, Bernado P, Blackledge M, and Sancho J. ProtSA: a web application for calculating sequence specific protein solvent accessibilities in the unfolded ensemble. *BMC Bioinformatics* 10: 104, 2009.
22. Fandrich M, Fletcher MA, and Dobson CM. Amyloid fibrils from muscle myoglobin. *Nature* 410: 165–166, 2001.
23. Fernandez-Busquets X, de Groot NS, Fernandez D, and Ventura S. Recent structural and computational insights into conformational diseases. *Curr Med Chem* 15: 1336–1349, 2008.
24. Foss TR, Wiseman RL, and Kelly JW. The pathway by which the tetrameric protein transthyretin dissociates. *Biochemistry* 44: 15525–15533, 2005.
25. Goldenberg DP and Creighton TE. Folding pathway of a circular form of bovine pancreatic trypsin inhibitor. *J Mol Biol* 179: 527–545, 1984.
26. Grantcharova VP and Baker D. Circularization changes the folding transition state of the src SH3 domain. *J Mol Biol* 306: 555–563, 2001.
27. Grantcharova VP, Riddle DS, and Baker D. Long-range order in the src SH3-folding transition state. *Proc Natl Acad Sci U S A* 97: 7084–7089, 2000.
28. Guijarro JI, Sunde M, Jones JA, Campbell ID, and Dobson CM. Amyloid fibril formation by an SH3 domain. *Proc Natl Acad Sci U S A* 95: 4224–4228, 1998.
29. Hagihara Y, Mine S, and Uegaki K. Stabilization of an immunoglobulin fold domain by an engineered disulfide bond at the buried hydrophobic region. *J Biol Chem* 282: 36489–36495, 2007.
30. Jahn TR and Radford SE. Folding versus aggregation: polypeptide conformations on competing pathways. *Arch Biochem Biophys* 469: 100–117, 2008.
31. Johnson CM, Oliveberg M, Clarke J, and Fersht AR. Thermodynamics of denaturation of mutants of barnase with disulfide crosslinks. *J Mol Biol* 268: 198–208, 1997.
32. Krittanai C and Johnson WC. Correcting the circular dichroism spectra of peptides for contributions of absorbing side chains. *Anal Biochem* 253: 57–64, 1997.
33. Liang J, Chen JK, Schreiber ST, and Clardy J. Crystal structure of P13K SH3 domain at 20 angstroms resolution. *J Mol Biol* 257: 632–643, 1996.
34. Martinez JC and Serrano L. The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nat Struct Biol* 6: 1010–1016, 1999.
35. Maurer-Stroh S, Debulpaep M, Kuemmerer N, Lopez de la Paz M, Martins IC, Reumers J, Morris KL, Copland A, Serpell L, Serrano L, Schymkowitz JW, and Rousseau F. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat Methods* 7: 237–242, 2010.
36. Mossuto MF, Bolognesi B, Guixer B, Dhulesia A, Agostini F, Kumita JR, Tartaglia GG, Dumoulin M, Dobson CM, and Salvatella X. Disulfide bonds reduce the toxicity of the amyloid fibrils formed by an extracellular protein. *Angew Chem Int Ed Engl* 50: 7048–7051, 2011.
37. Pace CN, Grimsley GR, Thomson JA, and Barnett BJ. Conformational stability and activity of ribonuclease T1 with zero, one, and two intact disulfide bonds. *J Biol Chem* 263: 11820–11825, 1988.
38. Parrini C, Bemporad F, Baroncelli A, Gianni S, Travaglini-Allocatelli C, Kohn JE, Ramazzotti M, Chiti F, and Taddei N. The folding process of acylphosphatase from *Escherichia coli* is remarkably accelerated by the presence of a disulfide bond. *J Mol Biol* 379: 1107–1118, 2008.
39. Riddle DS, Grantcharova VP, Santiago JV, Alm E, Ruczinski I, and Baker D. Experiment and theory highlight role of native state topology in SH3-folding. *Nat Struct Biol* 6: 1016–1024, 1999.
40. Sabate R, Castillo V, Espargaro A, Saupe SJ, and Ventura S. Energy barriers for HET-s prion forming domain amyloid formation. *FEBS J* 276: 5053–5064, 2009.
41. Sabate R, Espargaro A, de Groot NS, Valle-Delgado JJ, Fernandez-Busquets X, and Ventura S. 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, 2010.
42. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, and Serrano L. The FoldX web server: an online force field. *Nucleic Acids Res* 33: W382–W388, 2005.
43. Serio TR, Cashikar AG, Kowal AS, Sawicki GJ, Moslehi JJ, Serpell L, Arnsdorf MF, and Lindquist SL. Nucleated conformational conversion and the replication of conformational information by a prion determinant. *Science* 289: 1317–1321, 2000.
44. Stoll R and Bosserhoff A. Extracellular SH3 domain containing proteins—features of a new protein family. *Curr Protein Peptide Sci* 9: 221–226, 2008.
45. Strausberg S, Alexander P, Wang L, Gallagher T, Gilliland G, and Bryan P. An engineered disulfide cross-link accelerates the refolding rate of calcium-free subtilisin by 850-fold. *Biochemistry* 32: 10371–10377, 1993.
46. Ventura S, Lacroix E, and Serrano L. Insights into the origin of the tendency of the PI3-SH3 domain to form amyloid fibrils. *J Mol Biol* 322: 1147–1158, 2002.
47. Wani AH and Udgaonkar JB. Revealing a concealed intermediate that forms after the rate-limiting step of refolding of the SH3 domain of PI3 kinase. *J Mol Biol* 387: 348–362, 2009.
48. Zavodszky M, Chen CW, Huang JK, Zolkiewski M, Wen L, and Krishnamoorthi R. Disulfide bond effects on protein

stability: designed variants of Cucurbita maxima trypsin inhibitor-V. *Protein Sci* 10: 149–160, 2001.

49. Zhou HX. Effect of backbone cyclization on protein folding stability: chain entropies of both the unfolded and the folded states are restricted. *J Mol Biol* 332: 257–264, 2003.

Address correspondence to:
Prof. Salvador Ventura
Departament de Bioquímica i Biologia Molecular
Institut de Biotecnologia i de Biomedicina
Universitat Autònoma de Barcelona
Bellaterra 08193 (Barcelona)
Spain

E-mail: salvador.ventura@uab.es

Date of first submission to ARS Central, February 6, 2011; date of final revised submission, July 28, 2011; date of acceptance, July 28, 2011.

Abbreviations Used

ΔC_p =	heat capacity change of unfolding
ATR-FTIR =	attenuated total reflectance–Fourier transformed infrared spectroscopy
bis-ANS =	4,4′-dianilino-1,1′-binaphthyl-5,5′-disulfonic acid
CD =	circular dichroism
DSC =	differential scanning calorimetry
DTT =	dithiothreitol
Gdn·HCl =	guanidine hydrochloride
MALDI-TOF MS =	matrix-assisted laser desorption/ ionization-time of flight mass spectrometry
MIA =	melanoma inhibitory activity protein
PI3-SH3 =	SH3 domain of p85α subunit of bovine phosphatidyl-inositol-3′-kinase
SH3 =	SRC homology 3 domain
src-SH3 =	SH3 domain of src kinase
TCEP =	Tris (2-carboxyethyl)phosphine
Th-T =	thioflavin-T

4.3.- IDPRs acting as Entropic Bristles against Aggregation

The concept of entropic bristles was first introduced to describe a model for the maintenance of the intermolecular spacing between neurofilaments in the axon interior. According to this model, proteins with high flexibility, attached to the neurofilament core, would create a large excluded volume around them by experiencing large thermally-driven conformational fluctuations. This effect is considered to increase the effective volume of neurofilaments, thus helping its organization within the axon while still allowing the transit of small molecules (Brown & Hoh 1997). The idea of highly disordered protein extensions creating a excluded volume around individual molecules was later exploited to avoid the establishment of spurious intermolecular contacts leading to aggregation, thus allowing to increase the solubility of proteins where an engineered entropic bristle was translationally fused (Santner et al. 2012).

Proteins from the SUMO family are essential protein modifiers which all share the Ubiquitin-like (Ubl) fold. Aside from the structured Ubl domain, all the SUMO proteins whose structure has been resolved so far, from yeast to humans, possess an unstructured Nter polypeptide tail devoid of any obvious functional role. As well, the alignment of the sequences annotated as SUMO domains allows identifying the boundaries of the homologous Ubl region and reveals all of them bear a Nter extension. In the case of human SUMO2, this disordered extension harbors a consensus SUMOylation pattern that allows the formation of polySUMO chains. Protein modification with polySUMOs results in different modes of interaction from those arising by modification with monomeric SUMO (Xu et al. 2014). Nonetheless, the emergence of such a SUMOylation signature in the disordered tail of SUMO2 appears rather incidental, despite it may have been profited by the cell afterwards.

The inspection of the NMR conformers of the resolved solution structures of human SUMO domains shows how their disordered Nter tails create an excluded volume on top of a specific region of the Ubl structured domain. This particular region corresponds to the most amyloidogenic stretch of the SUMO domain, which is substantially exposed and coincides with the protein-protein interaction surface where SUMO interacting motifs (SIMs) bind to the globular domain (Namanja et al. 2012; Gareau et al. 2012). The presence of such an exposed amyloidogenic region substantially increases the risk of aggregation of the SUMO domains. Conversely, these proteins have been traditionally considered as a paradigm of protein solubility and have also been employed as fusion tags in order to improve the expression yield in heterologous protein production (Zuo et al. 2005; Marblestone et al. 2006). In a previous study we have shown that thermal destabilization of the SUMO native conformation results in the aggregation of these domains into amyloid-like fibrilar structures, whose core includes the mentioned amyloidogenic stretch. These observations indicate the native state protects the SUMO domain from aberrant aggregation despite possessing a potent APR that is significantly exposed to the solvent. At the same time, this suggests that exposure of the amyloidogenic stretch due to a perturbation of the native structure, or even during biosynthesis -since this APR is located in the Nter region of the domain- may lead to deleterious deposition, so the Nter unstructured tail may act as an entropic bristle by disfavoring the establishment of non-functional intermolecular interactions between SUMO polypeptides.

In order to test this hypothesis, the depositional behavior of SUMO domains, with and without its Nter unstructured tail, was evaluated. As a multitude of previous structural studies of the SUMO proteins indicated, the absence of the tail does not alter the conformation of the native state; additionally, it does not have any influence on the compactness of the Ubl domain, nor on its thermal stability. In contrast, the absence of the tail increases the exposure of hydrophobic patches on the surface of the protein, which is consistent with the idea of the Nter extension shielding the exposed amyloidogenic stretch. It also affects the hydrodynamic properties of the domain, since gel filtration analysis yields an apparent molecular mass difference between the full-length and the Nter deletion variant of SUMO2 around 7 times larger than the actual difference in mass between both forms, thus confirming that the high flexibility of the SUMO tail creates a large excluded volume. Finally, the Nter extension influences strongly the aggregation kinetics of the SUMO domain into amyloid-like fibrils upon thermal destabilization, while the variant devoid of the tail aggregates readily without any detectable lag phase, the full-length domain follows a kinetics consistent with the NCC mechanism, possessing an extended lag phase that last about the same time required for the aggregation of the tail-deleted variant to reach completion.

These findings provide strong evidence to consider the main role of the Nter disordered tails in SUMO proteins is functioning as entropic bristle domains. This is further supported by their sequential properties allowing to classify these extensions compositionally as IDPRs because of their low hydrophobicity, negative mean charge, and its depletion in “order-promoting” residues combined with enrichment in “disorder-promoting” ones. And also because of the ability of the SUMO tail to largely disrupt the aggregation of the Aβ42 peptide when translationally fused to its terminus.

Disordered regions are frequently found at the termini of proteins with a defined three-dimensional structure (Lobanov et al. 2010). Although terminal IDPRs serve to develop multiple functions (Uversky 2013b), the properties observed for the Nter tails of the SUMO domains suggest that acting as antiaggregational EB might be a widespread function of disordered termini. Indeed, the analysis of the patterns commonly found at disordered termini (Lobanov et al. 2010) shows they present a strikingly low tendency to aggregate, compared to the ensemble of globular proteins. This suggests that enrichment in “disorder-promoting” residues at the terminal ends of structured proteins may have evolved as a strategy to confront the establishment of spurious intermolecular contacts. This effect might be particularly relevant for proteins presenting potent APRs, which cannot be eliminated due to functional constraints in their Nter regions, as in the case of the SIM interaction site of SUMO domains, since during biosynthesis these regions might remain exposed for a considerable amount of time. Thus, evolving Nter disordered fragments with antiaggregational properties can help to avoid aberrant aggregation at this stage. A more specific analysis of functional constraints limiting the selection against aggregation prone regions is addressed in the following chapter.

Biomacromolecules

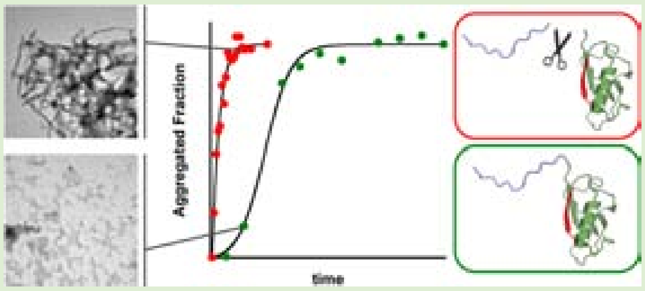
N-Terminal Protein Tails Act as Aggregation Protective Entropic Bristles: The SUMO Case

Ricardo Graña-Montes,[†] Patrizia Marinelli,[†] David Reverter, and Salvador Ventura*

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

S *Supporting Information*

ABSTRACT: The formation of β-sheet enriched amyloid fibrils constitutes the hallmark of many diseases but is also an intrinsic property of polypeptide chains in general, because the formation of compact globular proteins comes at the expense of an inherent sequential aggregation propensity. In this context, identification of strategies that enable proteins to remain functional and soluble in the cell has become a central issue in chemical biology. We show here, using human SUMO proteins as a model system, that the recurrent presence of disordered tails flanking globular domains might constitute yet another of these protective strategies. These short, disordered, and highly soluble protein segments would act as intramolecular entropic bristles, reducing the overall protein intrinsic aggregation propensity and favoring thus the attainment and maintenance of functional conformations.



INTRODUCTION

Protein misfolding and aggregation into β-sheet enriched amyloid-like structures are associated with a large set of human disorders, including Alzheimer's disease, diabetes, and some types of cancer.[1−4] However, the adoption of cytotoxic amyloid-like conformations is not restricted to disease-linked proteins and seems to constitute a generic property of polypeptide chains,[5,6] likely because the noncovalent contacts that stabilize native structures resemble those leading to the formation of amyloids.[7] Indeed, a majority of proteins contain at least one and often several aggregation-promoting sequences,[8,9] in many cases buried in the hydrophobic core of the native structure. Therefore, productive protein folding and deleterious aggregation are continuously competing in the cell. Because the formation of compact globular proteins comes at the expense of an inherent sequential aggregation propensity, identification of the strategies that enable proteins to remain functional and soluble in the cell is a central issue in biology.[10]

Organisms have evolved different mechanisms to survey and minimize side aggregation reactions,[11−15] including sophisticated and highly conserved protein quality control machineries.[16,17] In addition, during the course of evolution, proteins have adopted negative design strategies to prevent or diminish their intrinsic propensity to aggregate, by incorporating β-sheet breakers at structurally critical positions,[18,19] avoiding the presence of β-strands on the edge of protein structures[20] or placing gatekeeper residues at the flanks of aggregation-prone segments.[8] It has been recently suggested that the recurrent presence of disordered segments adjacent to folded domains in proteomes might be yet another strategy evolved to overcome aggregation.[21] Random movements of these tails around the point of attachment to the folded domain would sweep out a large area in space, acting thus as entropic bristles (EB).[22,23] Recently, the ability of long and highly disordered tails, either natural or artificial, to act as EB was tested experimentally by fusing them to different target proteins and expressing the fusions recombinantly in bacteria. Proteins fused to these EB were significantly more soluble than their natural counterparts.[22] Several evidences indicate that disordered terminal tails might also play an antiaggregational effect in the context of natural proteins. In this way, it has been shown that the disordered C-terminal region of NEIL1, a human homologue of *Escherichia coli* DNA glycosylase endonuclease VIII, is necessary for its soluble recombinant expression.[24] A similar role has been proposed for the highly charged and disordered C-terminal tail of α-synuclein in the context of α-synuclein-GST fusions.[25]

SUMO belongs to the ubiquitin-like (Ubl) protein family. The members of the Ubl family are small size (<10 KDa) post-translational modifiers, which are attached to protein substrates via an isopeptide bond between a C-terminal glycine and an acceptor lysine residue of the substrate.[26] Despite the low degree of sequence homology displayed between the members of the Ubl family, they all share a common protein fold and a similar mechanism of conjugation.[27,28] However, in contrast to ubiquitin, SUMO proteins contain an N-terminal extension of 15−20 amino acid residues, which constitutes a flexible tail that protrudes from the core protein.[29] Using N-terminal deletion mutants that include only the globular functional domain of

Received: December 4, 2013
Revised: February 6, 2014
Published: February 24, 2014

SUMO, we have previously shown that when the conformational stability of the SUMO1, SUMO2, and SUMO3 human isoforms is compromised, these small globular proteins aggregate into amyloid-like structures.[30] SUMO aggregation might have important physiological implications because the SUMO pathway is essential in mammals and in budding yeast. Therefore, it is likely that these proteins might have evolved strategies to minimize their aggregation propensity. By comparing the conformational, thermodynamic and aggregational properties of SUMO variants with and without the N-terminal extension, we show here that this tail acts as an EB, slowing down the aggregation kinetics of the globular domain. Moreover, by fusing the SUMO N-terminal tail to the highly amyloidogenic Aβ42 peptide and monitoring its impact on intracellular aggregation in bacteria,[31] we demonstrate that the solubilizing effect of this N-terminal extension can act in trans. We provide here experimental evidence for the antiaggregational effect of EB in a natural protein context. Computational analysis indicates that in fact this might be a generic function of disordered N- and C-terminal extensions in globular proteins.

■ MATERIALS AND METHODS

SUMO Domains Expression and Purification. Human SUMO2 full-length and a 14-residue Nter deletion variant (residues 15−95), referred here as SUMO2 and ΔNt-SUMO2, and SUMO1 full-length and a 17-residue Nter deletion variant (residues 18−101), denoted SUMO1 and ΔNt-SUMO1, respectively, were cloned into a pET-28b vector to encode either a Cter SENP2-cleavable hexahistidine fusion protein for full-length domains, or a Nter thrombin-cleavable hexahistidine fusion for Nter deletion variants. Cultures of *E. coli* BL21(DE3) cells transformed with these plasmids were grown in lysogeny broth (LB) medium with 50 μg·mL^{-1} kanamycin at 37 °C and 250 rpm to an OD at 600 nm of 0.5−0.6 before induction with 1 mM isopropyl-1-thio-β-D-galactopyranoside (IPTG) for 4 h at 30 °C. Next, the cultures were centrifuged and the cell pellets were frozen at −20 °C. After cell lysis, SUMO proteins were purified under native conditions by affinity chromatography on a FF-Histrap histidine-tag resin (General Electric). To cleave the histidine-tag, SUMO domains were incubated for 16 h at 4 °C with either SENP2 protease or thrombin, accordingly. The cleaved tags were removed by gel filtration on a HiLoad Superdex 75 prep grade column (General Electric). Protein buffer was further exchanged with the appropriate assay buffer on a Sephadex G-25 (General Electric) column prior to storage of SUMO samples at −80 °C. Unless otherwise stated, assays on SUMO domains were performed on 50 mM phosphate buffer at pH 7.

Intrinsic Fluorescence. SUMO2 and ΔNt-SUMO2 intrinsic fluorescence was monitored by recording Tyr emission spectra between 280 and 400 nm upon excitation at 268 nm. Spectra were registered, after equilibration at 298 K of a 50 μM protein sample, as the accumulation of three consecutive scans in a Jasco FP-8200 spectrofluorimeter (Jasco).

Intrinsic Fluorescence Quenching Assays. Quenching of SUMO2 and ΔNt-SUMO2 intrinsic fluorescence was analyzed by monitoring Tyr emission in the presence of acrylamide. Tyr fluorescent emission was recorded between 280 and 400 nm upon excitation at 268 nm, and after equilibration of 50 μM protein samples with final quencher concentrations ranging from 0 to 0.21 M at 298 K. Spectra were registered as the accumulation of three consecutive scans in a Jasco FP-8200 spectrofluorimeter (Jasco).

Bis-ANS Binding. Binding of 4,4'-bis(1-anilinonaphtalene 8-sulfonate) dye to soluble and aggregated forms of SUMO2 and ΔNt-SUMO2 was followed by recording the fluorescence spectra of protein−dye mixtures between 400 and 600 nm after excitation at 370 nm. Spectra were registered, after equilibration of the sample at 298 K, as the accumulation of three consecutive scans in a Jasco FP-8200 spectrofluorimeter (Jasco). Final protein and bis-ANS concentrations in the mixtures were 50 and 25 μM, respectively. Aggregated SUMO domains correspond to samples incubated at 343 K for 400 min.

Circular Dichroism (CD). SUMO2 and ΔNt-SUMO2 far-UV CD spectra were recorded for their soluble and aggregated forms between 195 and 270 nm at 298 K with a spectral resolution of 1 nm, using a Jasco 810 spectropolarimeter (Jasco). For spectra acquisition, protein samples were placed in a 0.1 cm path-length quartz cell at a final concentration of 20 μM and 20 scans were averaged for each spectrum.

Near-UV CD spectra were obtained for SUMO2 and ΔNt-SUMO2 soluble forms between 250 and 320 nm at 298 K with a spectral resolution of 1 nm. Spectra were obtained by accumulating 50 scans of 100 μM protein samples in a 0.2 cm cell, employing a Jasco 810 spectropolarimeter (Jasco).

Thermal Denaturation. SUMO2 and ΔNt-SUMO2 thermal denaturation was monitored by following the change of its CD signal at 222 nm and of its intrinsic fluorescence at 305 nm upon excitation at 268 nm. Signal change was recorded between 288 and 368 K using a 1 K·min^{-1} gradient at a final protein concentration of 20 μM. Experimental data were fitted to a two-state transition model, whose signals for the folded and unfolded states are dependent on the temperature, using the nonlinear least-squares algorithm provided with KaleidaGraph (Synergy Software).

Gel Filtration Chromatography. Gel filtration chromatography was employed routinely to remove purification tag peptides cleaved from SUMO domains, but it also allowed to infer differences between the effective hydrodynamic volumes of SUMO2 and ΔNt -SUMO2. For this analysis, ≈0.5 μmoles of previously purified SUMO2 or ΔNt -SUMO2 were injected in a HiLoad Superdex 75 prep grade column (General Electric) and eluted with 20 mM Tris, 250 mM NaCl, and 1 mM β-mercaptoethanol at pH 8.

Thioflavin-T (Th-T) Binding. Binding of the Th-T amyloid dye to SUMO2 and ΔNt-SUMO2 samples aggregated by incubating them for 400 min at 343 K was evaluated by recording the mixtures fluorescence spectra. Aggregated samples were diluted to 50 μM in Th-T, resulting in a final dye concentration of 25 μM. Spectra were recorded between 460 and 600 nm, after sample equilibration at 298 K and upon excitation at 440 nm, as the accumulation of three consecutive scans in a Jasco FP-8200 spectrofluorimeter (Jasco).

Transmission Electron Microscopy (TEM). Protein morphology was evaluated under TEM for SUMO2 and ΔNt-SUMO2 samples incubated for 400 min at 343 K. These samples were diluted 10-fold in distilled water and negatively stained by placing 10 μL of the dilutions on carbon-coated copper grids and allowing its deposition for 5 min, after sample removal the grids were stained with 2% uranyl acetate which was further removed after 1 min deposition. The grids were imaged in a Hitachi H-7000 transmission electron microscope operating at a 75 kV accelerating voltage.

Aggregation Kinetics. To follow its aggregation kinetics, soluble SUMO domains were prepared at a final protein concentration of 200 μM. In order to avoid the presence of preformed oligomeric species, protein samples were previously filtered through a 0.22 μm polyvinylidene fluoride filter. SUMO aggregation was triggered by incubating the proteins under 500 rpm agitation at temperatures close to their T_m: 343 K for SUMO2 and ΔNt-SUMO2, or 333 K for SUMO1 and ΔNt-SUMO1. Aggregation kinetics was monitored by evaluating the binding to the amyloid dye Thioflavin-T (Th-T) of samples from each aggregation reaction taken at different time intervals. Samples corresponding to each time point were diluted in Th-T and the fluorescence signal at 475 nm of the mixtures was recorded at 298K upon excitation at 440 nm in a Jasco FP-8200 spectrofluorimeter (Jasco). Final protein and dye concentrations in the mixtures were 100 and 25 μM, respectively.

The aggregated fraction of the protein (f) along the kinetics was calculated by normalizing Th-T fluorescence relative to the initial and end point intensities. Kinetic data from the full-length domains could be accuratelly fitted to an autocatalytic reaction model[32] according to the following equation:

Table 1. Physical, Intrinsic Disorder, and Solubility Properties Predicted for the N-ter Fragments of SUMO Family Members

								WH solubility[c] (%)		
protein	UniProt accession No.	net charge	pI	length (aa)	VLXT[a] (%)	VSL2[a] (%)	mean hydropathy[b]	full protein	ΔNter protein	Nter tail
Hs SUMO1	P63165	−3.1	4.42	18	67	100	−1.8	73.1	58.1	97.0
Hs SUMO2	P61956	−1.1	4.89	14	79	100	−1.7	64.6	61.4	80.3
Ce SUMO	P55853	−3.1	3.44	11	82	82	−0.8	86.8	73.1	97.0
At SUMO2	Q9FLP6	−2.1	4.42	14	93	100	−1.9	57.6	49.8	93.1
Sc Smt3	Q12306	−3.1	4.42	21	100	100	−1.4	81.1	72.5	97.7

[a]Percentage of amino acids found disordered using the PONDR predictors VLXT and VSL2. [b]A measure that distinguishes ordered from disordered proteins. [c]Solubility prediction according to the Wilkinson−Harrison method for the full-length protein, the protein without the Nter tail and the Nter fragment alone. The figures indicate the percentage of soluble protein.

$$f = \frac{\rho(e^{[(1+\rho)kt]} - 1)}{1 + \rho e^{[(1+\rho)kt]}} \quad (1)$$

where k is the product of the elongation constant of the aggregation reaction (k_e) times the protein concentration and ρ represents the dimensionless ratio of the nucleation constant (k_n) to k. k and ρ were derived by regression of f against time (t, in minutes) using the nonlinear least-squares algorithm provided with KaleidaGraph (Synergy Software). For aggregation kinetics with sigmoidal behavior the lag time was obtained by extrapolating the growth phase of the aggregation curve to $f = 0$, the half aggregation time and the time to aggregation completeness were derived using eq 1 and considering $f = 0.50$ or 0.99, respectively.

Analysis of Protein Solubility and Disorder in the SUMO Family. Theoretical analysis of the solubility and intrinsic disorder properties in the SUMO family was performed employing SUMO sequences from different species (Table 1). Protein solubility was estimated using the revised Wilkinson−Harrison solubility predictor,[33] and protein disorder was predicted with the PONDR-VLXT[34] and PONDR-VSL2[35] algorithms. All predictions were performed using the methods default settings.

Cloning and Expression of SUMO2-Aβ42-GFP Fusion. The insert encoding for human SUMO2 full-length was subcloned into a pET-28a vector already containing a fusion of Amyloid β 42 with the Green Fluorescent Protein (Aβ42-GFP), and was introduced upstream to the Aβ42 coding sequence. Next, a PCR was performed to amplify the whole plasmid but the region encoding for residues 15−95 of SUMO2 and the Cter hexahistidine tail, so the resulting product encodes for a ternary fusion of SUMO2 Nter tail (residues 1−14) with Aβ42-GFP (S2Nt-Aβ42-GFP). *E. coli* BL21(DE3) competent cells were transformed with the plasmid encoding for S2Nt-Aβ42-GFP or Aβ42-GFP without the SUMO2 Nter fragment. Bacterial cultures were grown at 37 °C and 250 rpm in LB medium containing 50 μg·mL^{-1} kanamycin and 34 μg·mL^{-1} chloramphenicol. When cultures reached an OD at 600 nm of 0.5, protein expression was induced with IPTG for 4 h. Cultures were further incubated for 16 h at 4 °C, then cells were pelleted and washed with phosphate-buffered saline (PBS) for three times.

GFP Fusions Fluorescence Determinations. GFP fluorescence in intact cells expressing the Aβ42-GFP or S2Nt-Aβ42-GFP fusions was measured at a cell density with an OD at 600 nm of 0.1, using a Jasco FP-8200 spectrofluorimeter (Jasco). GFP emission spectra were recorded between 500 and 600 nm at 298 K, after excitation at 470 nm.

Intact cells expressing GFP fusions were also imaged by phase-contrast microscopy and fluorescence microscopy under UV light using a Leica Q500 MC fluorescence DMBR microscope (Leica Microsystems) employing a 718 ms exposure. For microscopic analysis, 5 μL of washed cells were deposited on top of glass slides.

Aggregation Properties of Disordered Patterns at the Termini of Globular Proteins. Aggregation properties were analyzed for a set of protein fragments derived from the library of disordered patterns built by Galzitskaya and co-workers.[36] The sequences belonging to the library were filtered to exclude sequences with three or more consecutive His residues in order to avoid the presence of fragments which likely belong to affinity purification tags, thus having an artificial nature. The aggregation properties were analyzed for the resulting set, consisting of 71 sequences, using the AGGRESCAN[37] and Waltz[38] algorithms employing default settings. The AGGRESCAN aggregation parameters computed for the disordered patters were compared with those calculated for a reference set consisting of 71 sequences from the subset of sequences with less than 40% identity of the ASTRAL Compendium (Astral40), retrieved using a randomizing function.

RESULTS AND DISCUSSION

Conformational properties of SUMO2 and ΔNt-SUMO2. Full-length human SUMO2 is a 95 residues protein (Figure 1) in which the first 14 N-terminal residues correspond to a disordered tail according to its NMR 3D solution structure (PDB: 2AWT). This N-terminal extension is not involved in SUMO function, at least in vitro, and after its deletion, the remaining protein remains competent for the conjugation machinery.[39,40]

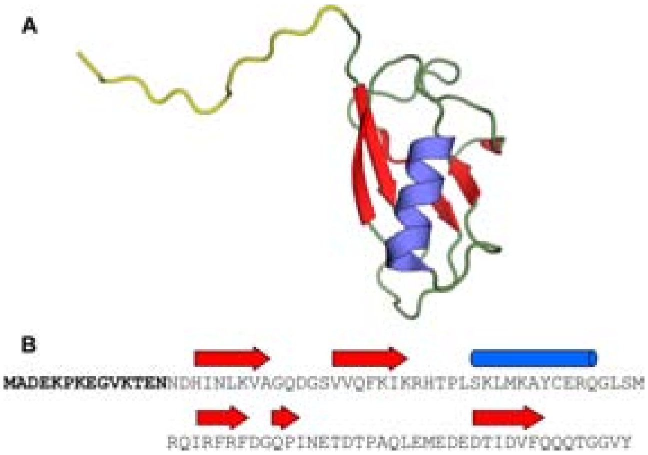

A
B
MADEKPKEGVKTENNDHINLKVAGQDGSVVQFKIKRHTPLSKLMKAYCERQGLSM
RQIRFRFDGQPINETDTPAQLEMEDEDTIDVFQQQTGGVY

Figure 1. Structure and sequence of SUMO2. (A) Cartoon representation of the full-length SUMO2 solution structure (PDB 2AWT). The N-terminal tail is shown in yellow. (B) Sequence and regular secondary structure elements of SUMO2. The N-terminal tail sequence is highlighted in bold.

We expressed and purified SUMO2 and a variant in which the first 14 N-terminal residues were deleted, thus, containing only the SUMO globular domain (ΔNt-SUMO2). We compared the secondary structure content of both protein variants using circular dichroism (CD) in the far-UV region (Figure 2A). Both spectra are essentially identical, deconvolution of the spectra using the K2D3 software[41] suggests that the β-sheet signal (34−45%) is the major contributor to the

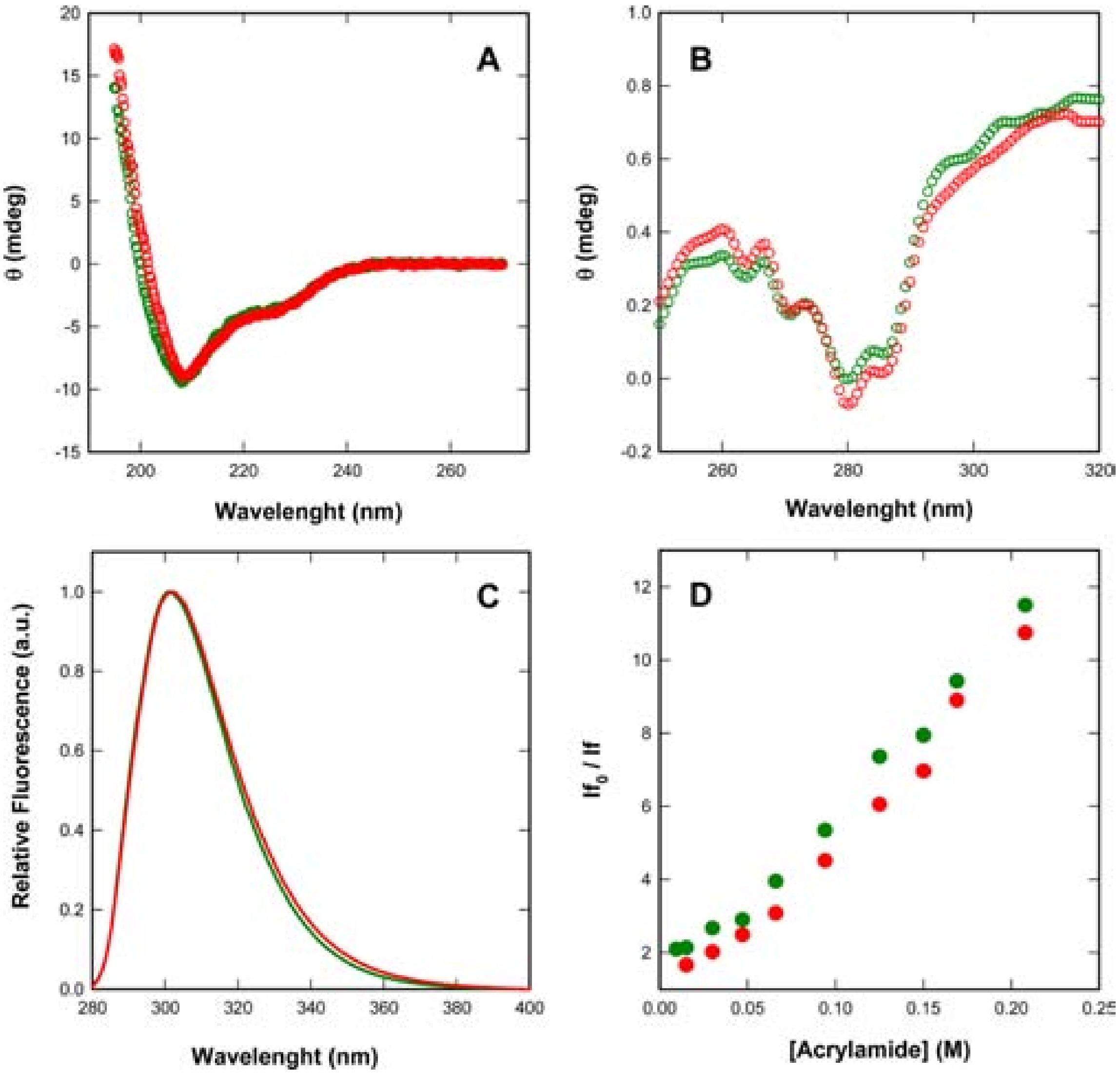


Figure 2. Conformational Properties of SUMO2 and ΔNt-SUMO2. (A) CD spectra in the far-UV region. (B) CD spectra in the near-UV region. (C) Intrinsic tyrosine fluorescence. (D) Stern−Volmer plot for the acrylamide quenching of tyrosine fluorescence. Data for SUMO2 and ΔNt-SUMO2 are shown in green and red, respectively.

spectra. We analyzed the CD spectra in the near-UV region to monitor differences in tertiary structure between the two variants. Also, in this region, the shapes of SUMO and ΔNt-SUMO2 spectra were almost identical (Figure 2B). Since SUMO2 lacks Trp residues we recorded the intrinsic fluorescence of Tyr residues in the two SUMO2 variants to monitor if they exhibit spectral differences. The proteins were excited at 268 nm and fluorescence recorded between 280 and 400 nm. The fluorescence spectra overlap, sharing the characteristic Tyr emission maximum at 305 nm (Figure 2C). Tyr fluorescence quenching by acrylamide was used to get more specific information on the location of these residues in both SUMO variants. As expected, in both cases, emission at 305 decreased with increasing acrylamide concentration. Stern−Volmer plots indicated that Tyr residues in the two proteins were in similar environments (Figure 2D). Overall, these data are in agreement with previous structural studies,[39] indicating that deletion of the N-terminal tail does not affect significantly the secondary and tertiary structure of the globular domain.

Thermal Stability of SUMO2 and ΔNt-SUMO2. The thermal stabilities of the two SUMO2 variants were analyzed by monitoring Tyr intrinsic fluorescence and far-UV CD changes at 305 and 222 nm, respectively, in the 293−368 K range. The transition curves from heat induced fluorescence emission changes in the SUMO2 variants are shown in Figure 3A. The thermal denaturation curves, followed by far-UV CD, are shown in Figure 3B. In both cases, a single cooperative transition was observed, and the data could be fitted to a two-state temperature-induced unfolding model ($R \geq 0.99$). The two probes reported essentially identical thermal transitions indicating that the secondary and tertiary structures are lost simultaneously upon heating, thus, supporting a two-state thermal unfolding mechanism for the two proteins.

The calculated melting temperatures (T_m) were 346.24 ± 0.35 and 345.06 ± 1.19 K, for SUMO2 and 345.65 ± 0.47 K and 344.08 ± 0.87 K for ΔNt-SUMO2, by intrinsic fluorescence and far-UV CD, respectively. These data indicate that deletion of the N-terminal tail does not affect the thermodynamic stability of SUMO2, which can be thus

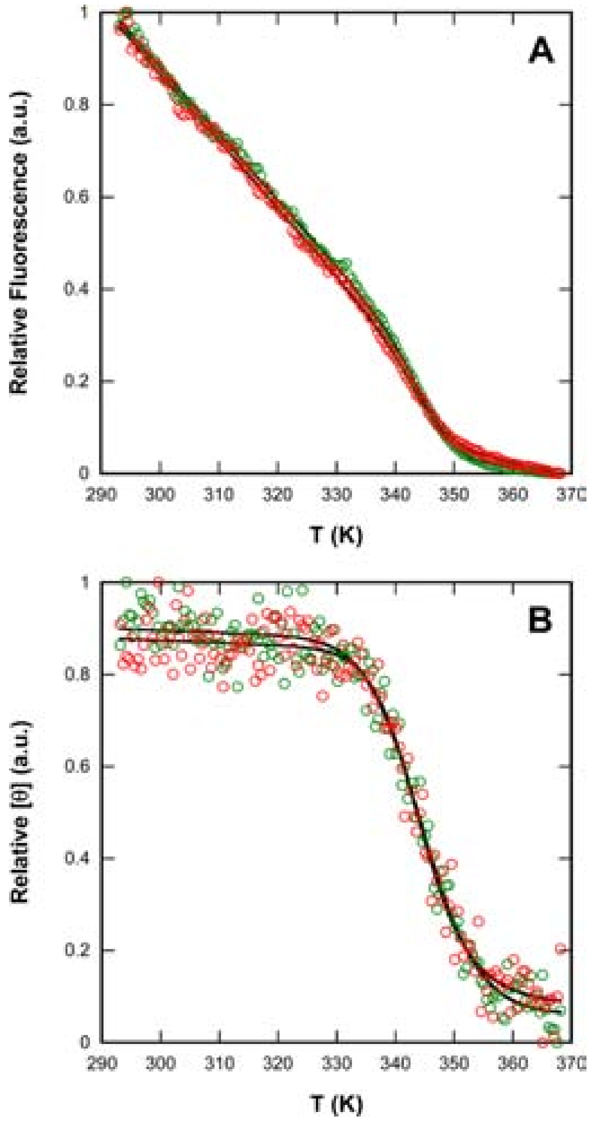


Figure 3. Thermal Denaturation of SUMO2 and ΔNt-SUMO2. Thermal denaturation was monitored by following changes in (A) intrinsic fluorescence and (B) CD signal at 222 nm of SUMO2 (green) and ΔNt-SUMO2 (red).

unequivocally attributed to the interactions sustaining the globular domain of the protein.

ΔNt-SUMO2 Exposes Hydrophobic Residues to Solvent. The N-terminal tail of SUMO2 might, in principle, affect the hydrodynamic properties of the full-length protein, when compared with that of ΔNt-SUMO2. We used gel-filtration chromatography to investigate if this is the case. As expected, SUMO2 elutes first than ΔNt-SUMO2 (Figure 4). However, the calculated difference in apparent molecular weight between the two proteins is ∼10 kDa, which is much higher than the one expected for a tail of only 14 residues (∼1.5 kDa). This suggests that the large difference in hydrodynamic volume between the two proteins can be attributed to the disordered nature of SUMO2 N-terminal extension.

Despite all previous data indicated that the globular domains in the two proteins were conformationally identical, the presence of a fluctuating tail might affect the domain surface properties. We monitored the presence of exposed hydrophobic clusters in the native structure of SUMO2 and ΔNt-SUMO2 by measuring their binding to 4,4′-bis(1-anilinonaphthalene 8-sulfonate) bis-ANS, a dye that increases its fluorescence emission upon interaction with these nonpolar regions.[42] We found that they display differential binding to this dye, ΔNt-SUMO2 exhibiting <2 times higher fluorescence emission than SUMO2 (Figure 5A). This implies that certain residues totally or partially protected from the solvent in SUMO2 increase their exposure after deletion of the N-terminal tail. According to the conformational and stability data this cannot be attributed to global or partial unfolding of the globular domain in the absence of the tail. Therefore, we speculated that the tail might shield, at least partially, exposed hydrophobic regions in the globular domain. An inspection of the 20 energy-minimized conformers in the NMR solution structure of the full-length protein indicates that this is likely the case (Figure 5B). Despite being devoid of any ordered motif, the N-terminus seems to dock preferentially on top of the globular domain, partially shielding β-strand 2, in which Val 29, Val30, and Phe32 are exposed to solvent.

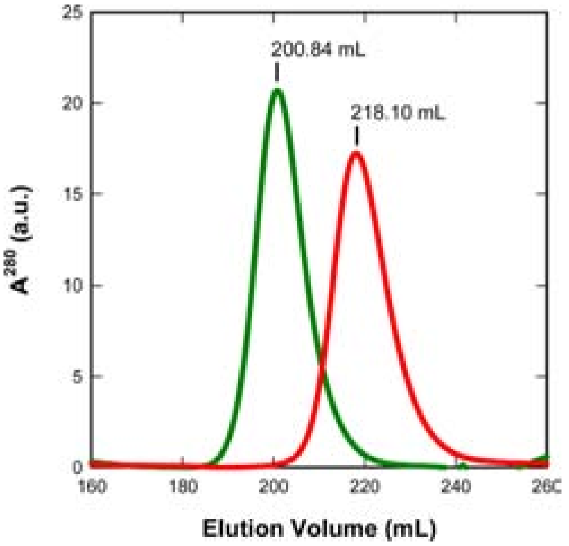


Figure 4. Hydrodynamic properties of SUMO2 and ΔNt-SUMO2. Gel filtration of SUMO2 (green) and ΔNt-SUMO2 (red) onto a Superdex 75 column. The elution volumes of the two proteins are indicated.

N-Terminal Tail Protects SUMO Proteins from Aggregation into Amyloid-Like Structures. We have shown in a recent study that β-strand 2 is the most amyloidogenic region in the sequence of SUMO2 and it is fully protected in the amyloid-fibrils formed by the globular domain under destabilizing conditions.[30] The observation that the N-terminal tail could partially shield this strand from the solvent might imply a protective role of the tail against the aggregation of the globular domain. To test this possibility we incubated SUMO2 and ΔNt-SUMO2 at 343 K for 400 min and afterward we monitored their binding to the amyloid detecting dye Thioflavin-T (Th-T; Figure 6A). Despite both samples promoted an increase in Th-T fluorescence emission, the presence of the N-terminal tail reduced Th-T fluorescence emission by 5-fold in comparison to the globular domain alone. Analysis of the two incubated protein solutions by far-UV CD after cooling at 293 K (Figure 6B), indicates that SUMO2 maintains/recovers essentially its native spectrum in terms of shape and intensity. In contrast, in the case of ΔNt-SUMO2, despite the spectrum is essentially native, the intensity is 3-fold

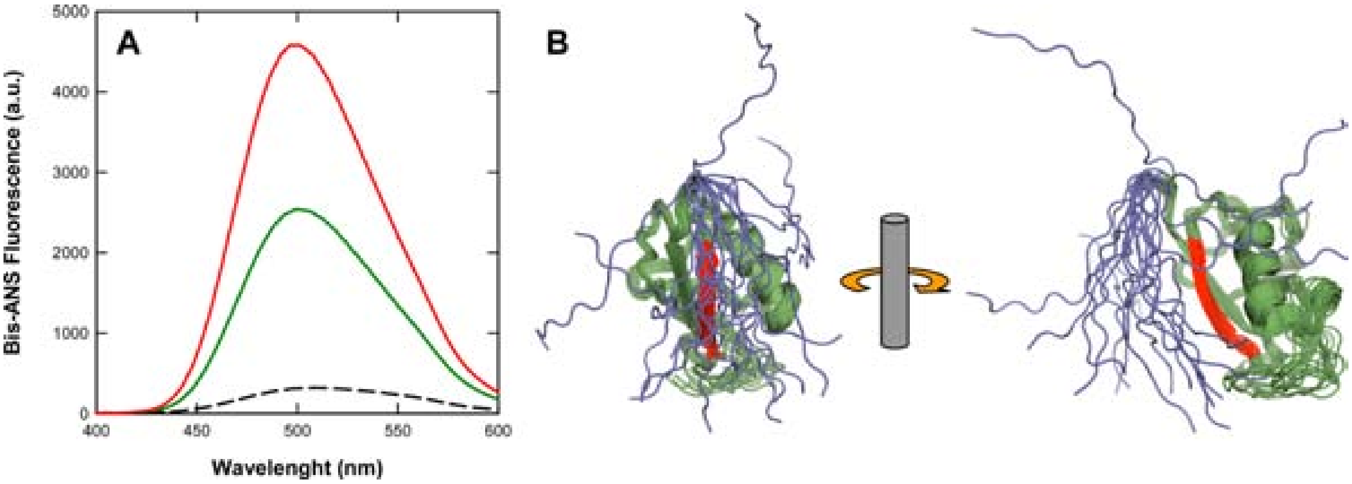


Figure 5. Exposure of Hydrophobic Clusters in SUMO2 and ΔNt-SUMO2. (A) bis-ANS binding to SUMO2 (green) and ΔNt-SUMO2 (red). The slashed line represents free bis-ANS emission spectrum. (B) Cartoon representations of the 20 conformers with lowest energy derived from the full-length (residues 1−95) SUMO2 solution structure determination (PDB: 2AWT). The conformers are structurally aligned considering solely the structured globular ubiquitin-like domain (residues 15−93). The Nter unstructured tail is shown in blue, and a previously identified aggregation-prone region [27] in red. The representations are rotated 90° over the z-axis one respect the other.

lower than in the case of SUMO2, suggesting that, in agreement with the Th-T binding data, a significant fraction of the protein has precipitated out of the solution. We used transmission electron microscopy to confirm these data. As it can be seen in Figure 6C, the ΔNt-SUMO2 solution exhibits abundant amyloid fibrils, whereas the presence of the N-terminal tail completely abrogated the presence of long fibrils and only small aggregates were detected in SUMO2 solutions.

We addressed if the observed differences in aggregation properties might have a kinetic origin by monitoring the time course of SUMO2 and ΔNt-SUMO2 aggregation at 343 K using Th-T (Figure 7A). The kinetics of ΔNt-SUMO2 aggregation is fast, following an hyperbolic curve, reaching a plateau after 300 min and lacking any detectable lag phase. In contrast, SUMO2 exhibits a canonical sigmoidal aggregation curve that reflects a nucleation-dependent growth mechanism, with a lag time of 385 min, half aggregation time ($t_{1/2}$) of 824 min and reaching the final aggregated stated after 1747 min, being thus exceedingly slow when compared with the ΔNt-SUMO2 aggregation process.

As in the case of SUMO2, human SUMO1 possesses a N-terminal disordered tail consisting in this case of 18 residues (Figure 8). To assess if SUMO1 N-terminal tail plays also a protective role against aggregation, we expressed and purified SUMO1 and a N-terminal deletion mutant consisting only of the globular domain (ΔNt-SUMO1), we incubated these two proteins at their melting temperature (333 K) and followed their aggregation kinetics as explained above. As it can be observed in Figure 7B, also in the case of SUMO1, the presence of the N-terminal tail results in a significantly slower aggregation reaction.

SUMO N-Terminal Tails Resemble Entropic Bristles. The presence of N-terminal extensions is a recurrent feature of SUMO proteins in eukaryotic organisms, from yeast to humans (Figure 8).

Table 1 shows the characteristics of disorder and solubility predictions for the N-terminal tails of human SUMO1 and SUMO2 and the SUMO proteins from *C. elegans*, *S. cerevisiae*, and *A. taliana*. PONDR predictors VLXT and VSL2[34,35] confirm a high disorder propensity in all N-terminal extensions. In fact, despite the absence of sequence homology between SUMO N-terminal tails, they all share a net negative charge and a low mean hydrophobicity, a characteristic of disordered proteins. In terms of composition they only exhibit 7% of the so-called order-promoting amino acids (W, Y, F, I, L, V, C, and N), whereas 67% of the residues correspond to disorder-promoting amino acids (A, R, G, Q, S, E, K, and P). Using the Wilkinson−Harrison solubility model[33] we also show that all N-terminal tails are predicted to be highly soluble peptides and that their presence in full length SUMO proteins is predicted to increase the solubility of the corresponding globular domain in all cases. Overall, despite being shorter, N-terminal SUMO tails share many features with canonical EB, defined as long heterologous intrinsically disordered sequences able to enhance the solubility of aggregation-prone proteins when fused in trans at their N-terminus.[22]

SUMO2 N-Terminal Tail Reduces the Intracellular Aggregation Propensity of the Aβ42 Peptide. We investigated if as it happens with EB, the SUMO2 N-terminal tail can also act in trans by reducing the aggregation propensity of an insoluble polypeptide when fused at its N-terminal side. We used a fusion to the highly amyloidogenic Aβ42 peptide as a proof of principle. We have previously shown that, when fused to the green fluorescent protein (GFP) and expressed in *E. coli*, the Aβ42-GFP fusion forms inclusion bodies (IB) displaying amyloid-like properties.[43] Using a large set of mutants, we have demonstrated that, in this system, the presence of active GFP in the aggregates depends on the aggregation propensity of the Aβ42 variant.[44] The faster the fusion protein aggregates, the lower the IB fluorescence emission is and vice versa, in such a way that the fluorescence of IBs reports on the in vivo protein aggregation.[45]

We fused the N-terminal tail of SUMO2 at the N-terminus of Aβ42-GFP and expressed the new fusion (NterS2-Aβ42-GFP) in bacteria. The original Aβ42-GFP fusion was expressed as a control. Next, we measured the GFP fluorescence of intact cells expressing the two variants using spectrofluorimetry. As shown in Figure 9A, the two fusions differ in their fluorescence emission, cells expressing NterS2-Aβ42-GFP variant being three times more fluorescent than control cells. We used

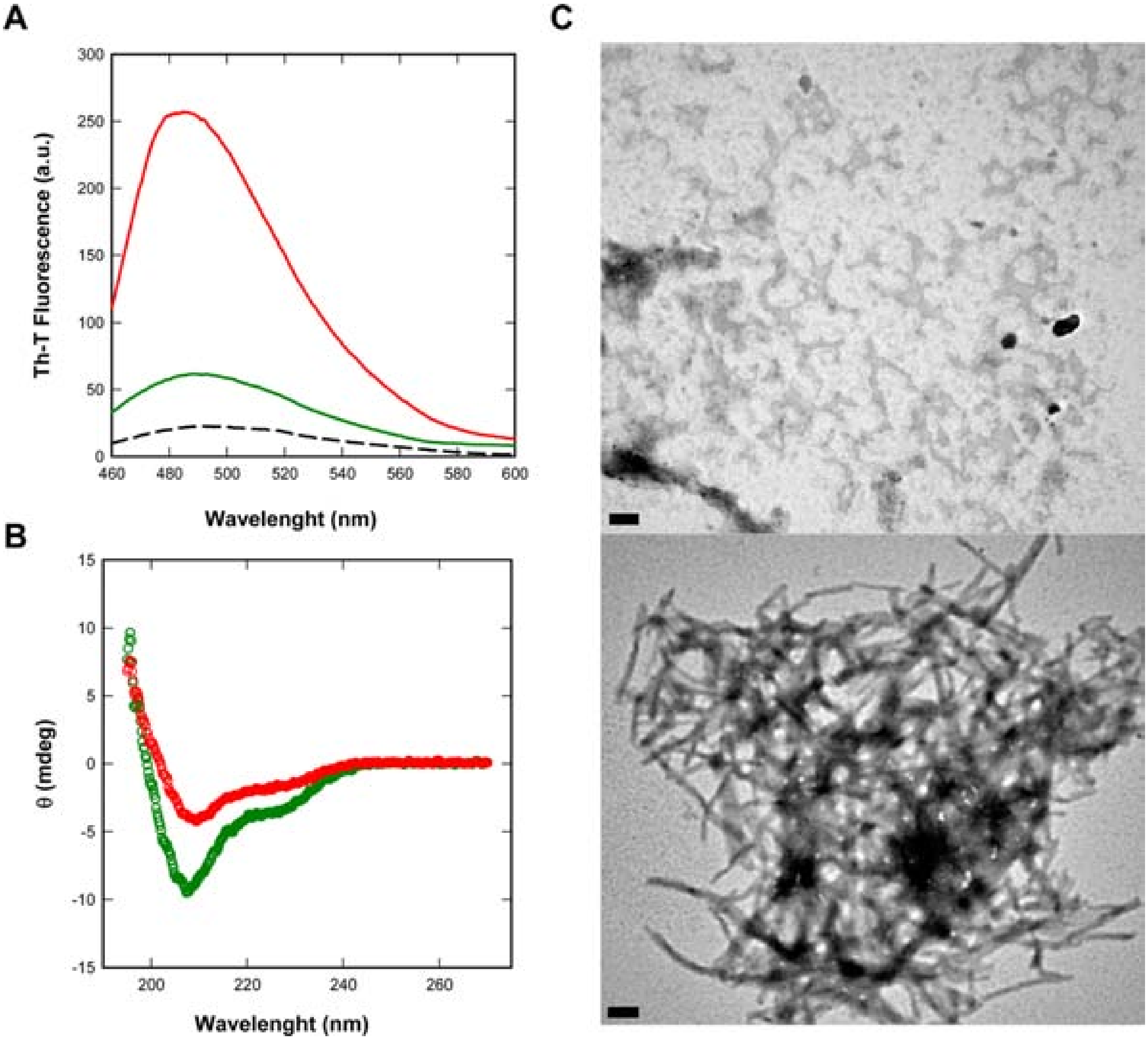


Figure 6. Morphological, Dye Binding, and Structural Properties of SUMO2 and ΔNt-SUMO2 Aggregates. (A) Th-T emission spectra and (B) CD spectra of SUMO2 (green) and ΔNt-SUMO2 (red) aggregates after 400 min incubation. The slashed line represents the emission spectrum of free Th-T. (C) Representative TEM micrographs showing aggregate morphology after 400 min incubation of SUMO2 (top) and ΔNt-SUMO2 (bottom). The scale bar represents 1 μm.

fluorescence microscopy to identify the cellular location of the detected GFP emission (Figure 9B). In both cases, the fluorescence is confined mainly in the IBs at the poles of the cell. However, the IBs formed by NterS2-Aβ42-GFP are clearly more fluorescent than those formed by Aβ42-GFP alone, indicating that the fusion in trans of the N-terminal tail of SUMO2 significantly reduces the intracellular aggregation propensity of the amyloidogenic Aβ42 peptide. This solubilizing effect resembles the one exerted by an artificial segment comprising 19 repeats of the tetrapeptide sequence NANP, when fused at the N-terminus of Aβ42 peptide.[46]

Disordered Patterns at the N- and C-Termini of Globular Domains Display Low Aggregation Propensity. We wondered whether the antiaggregational properties of the N-terminal SUMO tails might be in fact a generic property of short sequences flanking globular domains. To study this possibility we exploited a library of disordered patterns in 3D structures developed by Galzitskaya and co-workers.[36] They identified 109 disordered patterns of different lengths in disordered regions in the PDB. We examined the properties of the patterns identified at the first (N-tail) and last (C-tail) 40 residues of proteins. We analyzed their overall aggregation propensity and the presence of aggregation-prone regions using AGGRESCAN and their amyloidogenicity as predicted by WALTZ. We excluded from the analysis all the patterns containing three or more consecutive His residues, since they might correspond to artificial His-tags, reducing the data set to 71 different patterns (Table S1). AGGRESCAN indicates that the aggregation tendency of these sequences is extremely low, with an average value of −59.06, which contrasts with an average aggregation propensity of 0.023 for the ensemble of protein sequences in Swiss-Prot (the more negative the value

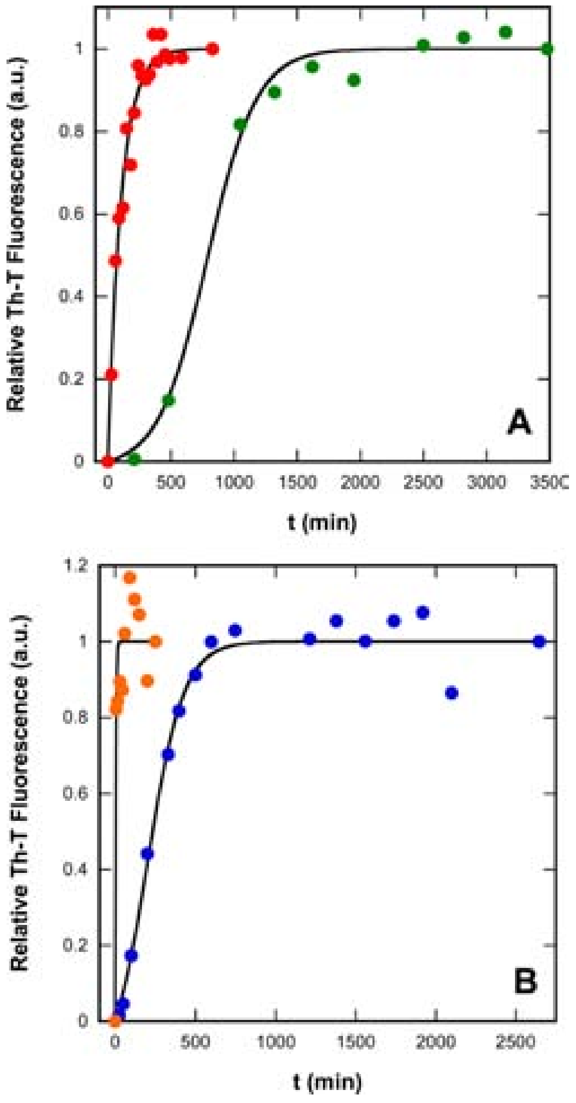


Figure 7. Aggregation Kinetics of SUMO Domains. The aggregation kinetics was monitored by following the change in relative Th-T fluorescence at different time points for (A) of SUMO2 (green) and ΔNt-SUMO2 (red), and (B) SUMO1 (blue) and ΔNt-SUMO1 (orange).

```
Hs SUMO1     MSDQEAKPSTEDLGDKKEGEYIKLKVIGQDSSEIHFKVK 39
Hs SUMO2         MADEKPKEGVKTENNDHINLKVAGQDGSVVQFKIK 35
Ce SUMO             MADDAAQAGDNAEYIKIKVVGQDSNEVHFRVK 32
At SUMO2          MSATPEEDKKPDQGAHINLKVKGQDGNEVFFRIK 34
Sc Smt3    MSDSEVNQEAKPEVKPEVKPETHINLKVS-DGSSEIFFKIK 40
                       .     .   :*::**  : .. : *::*
```

Figure 8. Aligment of SUMO Sequences from Different Species. SUMO Nter region sequences from *Homo sapiens* (Hs), *Caenorhabditis elegans* (Ce), *Arabidopsis thaliana* (At), and *Saccharomyces cerevisiae* (Sc). The highlighted segment denotes the fragment which does not belong to the ubiquitin-like fold and that is unstructured in Hs SUMO1 and SUMO2, according to their solution structures (PDBs 1A5R and 2AWT, respectively).

the more soluble the protein). Comparison of the distribution of aggregation propensities of these sequential patterns with those of the sequences of 71 proteins randomly selected from the SCOP-derived ASTRAL40 data set[47] illustrates the comparatively low aggregation tendency of protein termini (Figure 10). In fact, 97% of the patterns are predicted to be devoid of aggregation-prone or amyloidogenic sequences, thus, being potential EB.

■ CONCLUSIONS

It is now widely accepted that protein misfolding and aggregation impact cell physiology, reducing organisms cell fitness. Thus, selection against protein aggregation is expected to act as an important constraint in the evolution of protein sequences. Accordingly, proteins have adopted different structural and sequential strategies to prevent or reduce their aggregation propensity.

Disordered regions have been found in a large number of eukaryotic proteins.[48] These stretches are not equally distributed within protein sequences, with residues at the protein termini displaying on the average higher disorder propensity.[36] Disordered tails are not just flexible protrusions, instead they are being found associated with an increasing range of protein functions.[23] Computational simulations have suggested that one of the functions of disordered segments in proteins might be reduction of the aggregation propensity of the flanked domains.[21] In particular, multiscale simulations have shown that the aggregation of fungal hydrophobins at air—water interfaces relies on the reduction of the flexibility of an inner loop, which is disordered in aqueous solution, where these proteins remain soluble.[49] The characterization of the conformational and aggregative properties of SUMO and ΔNt-SUMO variants in the present study provides yet another experimental evidence for the antiaggregational role of the usually highly flexible termini in the context of a natural protein.

In vitro, the SUMO2 N-terminal tail does not affect significantly the secondary and tertiary structure of the globular domain, neither its conformational stability nor its function. However, it might transiently shield exposed hydrophobic residues in the amyloidogenic β-strand 2. This β-strand forms a functional intermolecular β-sheet with SUMO partners containing SUMO binding domains, being a major contributor to complex stability. Therefore, its intrinsic aggregation propensity responds to functional constraints and cannot be evolutively suppressed. The flexible N-terminal tail would contribute to decrease this aggregation tendency without compromising function.

The sequences of SUMOs N-tails in different organisms resemble the so-called EB. The solubilizing effects of disordered EB on globular proteins respond to two major effects: (1) a relatively larger surface able to interact favorably with water and (2) a large excluded volume that would restrict intermolecular contacts between aggregation-prone regions in proteins. These two factors are highly inter-related and would act increasing the entropic cost for protein aggregation resulting in a higher free energy barrier. In fact, the behavior of EB resembles that of polymers like polyethylene glycol,[22] which chemical attachment to globular proteins significantly increases their solubility.[50] This generic ability to increase solubility explains why these disordered segments, as we show here for SUMO2 N-terminal tail, can act in trans.[22] The computational analysis of the most frequent sequential patterns in disordered N- and C- termini shows that they are highly soluble and skip the presence of dangerous aggregation/amyloidogenic sequences, arguing that their recurrent presence in proteins might constitute a negative design strategy to maintain solubility. Our results suggest that dynamic aspects should be taken into account when addressing the aggregation/solubility properties of proteins.

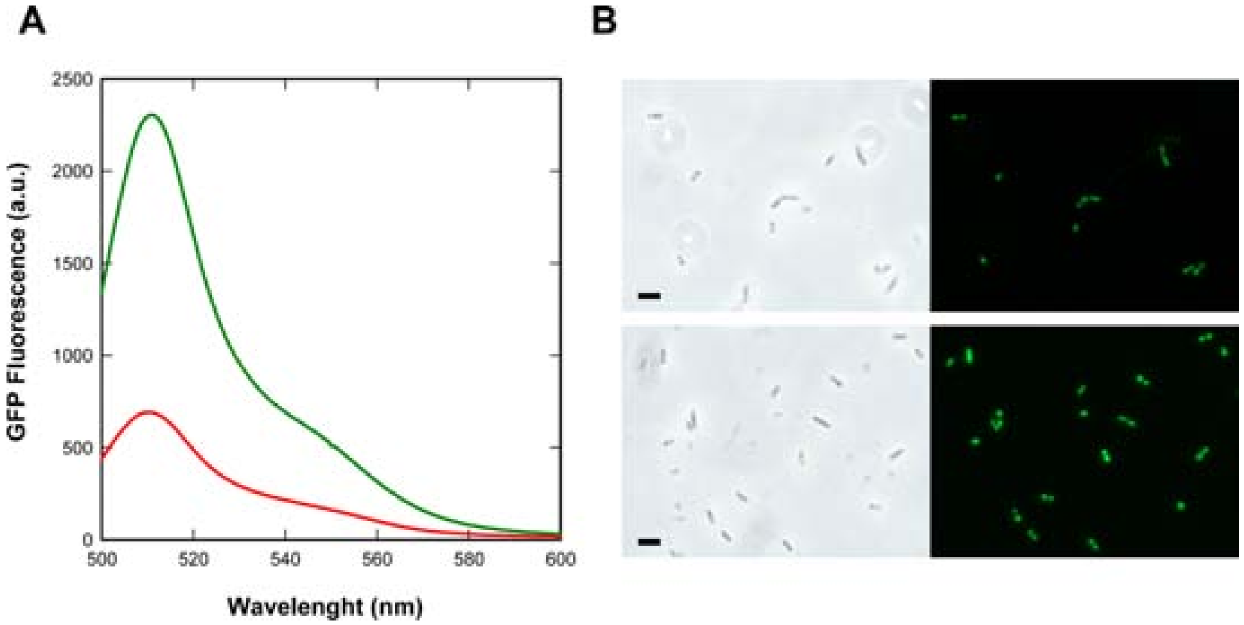


Figure 9. In vivo analysis of GFP fusions solubility. (A) Emission spectra of intact cells suspensions expressing the Aβ42-GFP fusion (red) or the S2Nt-Aβ42-GFP fusion (green). (B) Intact cells imaged using phase-contrast microscopy (left) and fluorescence microscopy (right). Upper panels show bacteria expressing the Aβ42-GFP fusion and lower panels cells expressing the S2Nt-Aβ42-GFP fusion. The scale bar represents 5 μm.

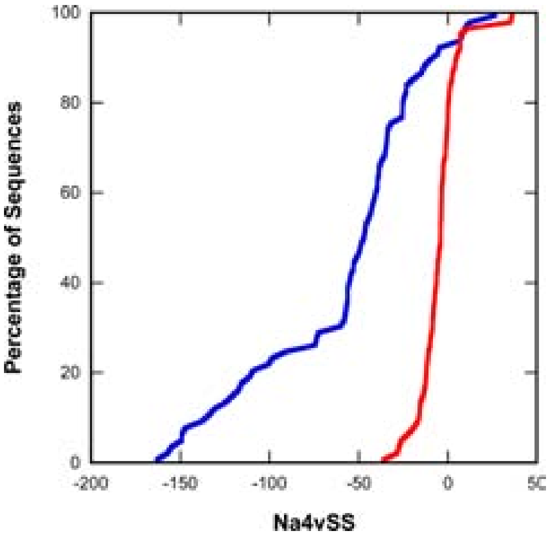


Figure 10. Aggregation properties of disordered patterns and globular proteins. Distribution of the average aggregation propensity (Na4vSS) computed by the AGGRESCAN algorithm for 71 disordered patterns (blue) retrieved from Galzitskaya and co-workers library[36]and 71 globular proteins (red) randomly selected from the Astral40 data set.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Table S1. List of the 71 patterns retrieved from the disordered patterns library constructed by Galzitskaya and co-workers.[36] This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Tel.: 34-93-5868956. Fax: 34-93-5811264. E-mail: salvador.ventura@uab.es.

Author Contributions

[†]These authors contributed equally (R.G.-M. and P.M.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Work in our lab is supported by Grants BFU2010-14901 from Ministerio de Ciencia e Innovacion (Spain) and 2009-SGR-760 from AGAUR (Generalitat de Catalunya). S.V. has been granted an ICREA Academia award (ICREA). R.G.-M. is the beneficiary of the FPU AP2009-0948 fellowship from the Ministerio de Educación (Spain) and P.M. is the beneficiary of a FI-DGR 2012 fellowship from the AGAUR (Generalitat de Catalunya).

■ REFERENCES

(1) Chiti, F.; Dobson, C. M. *Annu. Rev. Biochem.* **2006**, *75*, 333.

(2) Fernandez-Busquets, X.; de Groot, N. S.; Fernandez, D.; Ventura, S. *Curr. Med. Chem.* **2008**, *15*, 1336.

(3) Silva, J. L.; Rangel, L. P.; Costa, D. C.; Cordeiro, Y.; De Moura Gallo, C. V. *Biosci. Rep.* **2013**, *33*, e00054.

(4) Xu, J.; Reumers, J.; Couceiro, J. R.; De Smet, F.; Gallardo, R.; Rudyak, S.; Cornelis, A.; Rozenski, J.; Zwolinska, A.; Marine, J. C.; Lambrechts, D.; Suh, Y. A.; Rousseau, F.; Schymkowitz, J. *Nat. Chem. Biol.* **2011**, *7*, 285.

(5) Jahn, T. R.; Radford, S. E. *FEBS J.* **2005**, *272*, 5962.

(6) Dobson, C. M. *Semin. Cell Dev. Biol.* **2004**, *15*, 3.

(7) Linding, R.; Schymkowitz, J.; Rousseau, F.; Diella, F.; Serrano, L. *J. Mol. Biol.* **2004**, *342*, 345.

(8) Rousseau, F.; Serrano, L.; Schymkowitz, J. W. *J. Mol. Biol.* **2006**, *355*, 1037.

(9) Goldschmidt, L.; Teng, P. K.; Riek, R.; Eisenberg, D. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 3487.

(10) Chiti, F.; Dobson, C. M. *Nat. Chem. Biol.* **2009**, *5*, 15.

(11) Gsponer, J.; Babu, M. M. *Cell Rep.* **2012**, *2*, 1425.

(12) Sanchez de Groot, N.; Torrent, M.; Villar-Pique, A.; Lang, B.; Ventura, S.; Gsponer, J.; Babu, M. M. *Biochem. Soc. Trans.* **2012**, *40*, 1032.

(13) Monsellier, E.; Chiti, F. *EMBO Rep.* **2007**, *8*, 737.

(14) Tartaglia, G. G.; Pechmann, S.; Dobson, C. M.; Vendruscolo, M. *Trends Biochem. Sci.* **2007**, *32*, 204.

(15) Reumers, J.; Maurer-Stroh, S.; Schymkowitz, J.; Rousseau, F. *Hum. Mutat.* **2009**, *30*, 431.

(16) Tyedmers, J.; Mogk, A.; Bukau, B. *Nat. Rev. Mol. Cell Biol.* **2010**, *11*, 777.

(17) Hartl, F. U.; Bracher, A.; Hayer-Hartl, M. *Nature* **2011**, *475*, 324.

(18) Steward, A.; Adhya, S.; Clarke, J. *J. Mol. Biol.* **2002**, *318*, 935.

(19) Parrini, C.; Taddei, N.; Ramazzotti, M.; Degl'Innocenti, D.; Ramponi, G.; Dobson, C. M.; Chiti, F. *Structure (Oxford, U. K.)* **2005**, *13*, 1143.

(20) Richardson, J. S.; Richardson, D. C. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 2754.

(21) Abeln, S.; Frenkel, D. *PLoS Comput. Biol.* **2008**, *4*, e1000241.

(22) Santner, A. A.; Croy, C. H.; Vasanwala, F. H.; Uversky, V. N.; Van, Y. Y.; Dunker, A. K. *Biochemistry* **2012**, *51*, 7250.

(23) Uversky, V. N. *FEBS Lett.* **2013**, *587*, 1891.

(24) Bandaru, V.; Cooper, W.; Wallace, S. S.; Doublie, S. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *60*, 1142.

(25) Park, S. M.; Jung, H. Y.; Chung, K. C.; Rhim, H.; Park, J. H.; Kim, J. *Biochemistry* **2002**, *41*, 4137.

(26) Johnson, E. S. *Annu. Rev. Biochem.* **2004**, *73*, 355.

(27) Hershko, A.; Ciechanover, A. *Annu. Rev. Biochem.* **1998**, *67*, 425.

(28) Hay, R. T. *Mol. Cell* **2005**, *18*, 1.

(29) Bayer, P.; Arndt, A.; Metzger, S.; Mahajan, R.; Melchior, F.; Jaenicke, R.; Becker, J. *J. Mol. Biol.* **1998**, *280*, 275.

(30) Sabate, R.; Espargaro, A.; Grana-Montes, R.; Reverter, D.; Ventura, S. *Biomacromolecules* **2012**, *13*, 1916.

(31) Villar-Pique, A.; de Groot, N. S.; Sabate, R.; Acebron, S. P.; Celaya, G.; Fernandez-Busquets, X.; Muga, A.; Ventura, S. *J. Mol. Biol.* **2012**, *421*, 270.

(32) Sabate, R.; Villar-Pique, A.; Espargaro, A.; Ventura, S. *Biomacromolecules* **2012**, *13*, 474.

(33) Wilkinson, D. L.; Harrison, R. G. *Bio/Technology* **1991**, *9*, 443.

(34) Romero, P.; Obradovic, Z.; Li, X.; Garner, E. C.; Brown, C. J.; Dunker, A. K. *Proteins* **2001**, *42*, 38.

(35) Peng, K.; Radivojac, P.; Vucetic, S.; Dunker, A. K.; Obradovic, Z. *BMC Bioinf.* **2006**, *7*, 208.

(36) Lobanov, M. Y.; Furletova, E. I.; Bogatyreva, N. S.; Roytberg, M. A.; Galzitskaya, O. V. *PLoS Comput. Biol.* **2010**, *6*, e1000958.

(37) Conchillo-Sole, O.; de Groot, N. S.; Aviles, F. X.; Vendrell, J.; Daura, X.; Ventura, S. *BMC Bioinf.* **2007**, *8*, 65.

(38) 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. *Nat. Methods* **2010**, *7*, 237.

(39) Reverter, D.; Lima, C. D. *Nature* **2005**, *435*, 687.

(40) Pichler, A.; Knipscheer, P.; Oberhofer, E.; van Dijk, W. J.; Korner, R.; Olsen, J. V.; Jentsch, S.; Melchior, F.; Sixma, T. K. *Nat. Struct. Mol. Biol.* **2005**, *12*, 264.

(41) Louis-Jeune, C.; Andrade-Navarro, M. A.; Perez-Iratxeta, C. *Proteins* **2011**, DOI: 10.1002/prot.23188.

(42) de Groot, N. S.; Parella, T.; Aviles, F. X.; Vendrell, J.; Ventura, S. *Biophys. J.* **2007**, *92*, 1732.

(43) Morell, M.; Bravo, R.; Espargaro, A.; Sisquella, X.; Aviles, F. X.; Fernandez-Busquets, X.; Ventura, S. *Biochim. Biophys. Acta* **2008**, *1783*, 1815.

(44) de Groot, N. S.; Aviles, F. X.; Vendrell, J.; Ventura, S. *FEBS J.* **2006**, *273*, 658.

(45) Villar-Pique, A.; de Groot, N. S.; Sabate, R.; Acebron, S. P.; Celaya, G.; Fernandez-Busquets, X.; Muga, A.; Ventura, S. *J. Mol. Biol.* **2012**, *421*, 270.

(46) Finder, V. H.; Vodopivec, I.; Nitsch, R. M.; Glockshuber, R. *J. Mol. Biol.* **2010**, *396*, 9.

(47) Chandonia, J. M.; Hon, G.; Walker, N. S.; Lo Conte, L.; Koehl, P.; Levitt, M.; Brenner, S. E. *Nucleic Acids Res.* **2004**, *32*, D189.

(48) Ward, J. J.; Sodhi, J. S.; McGuffin, L. J.; Buxton, B. F.; Jones, D. T. *J. Mol. Biol.* **2004**, *337*, 635.

(49) De Simone, A.; Kitchen, C.; Kwan, A. H.; Sunde, M.; Dobson, C. M.; Frenkel, D. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 6951.

(50) Milla, P.; Dosio, F.; Cattel, L. *Curr. Drug Metab.* **2012**, *13*, 105.

■ NOTE ADDED AFTER ASAP PUBLICATION

This article posted ASAP on March 6, 2014. Numerous changes have been made throughout the paper. The correct version posted March 10, 2014.

4.4.- Protein Abundance as a Modulator of Aggregative Risk in the Cellular Environment

It has already been introduced that aggregation constitutes a high-order reaction, which is highly dependent on protein concentration; therefore regulation of protein abundance appears to be a primary mechanism the cell exploits in order to manage the risk of aggregation. The control of protein concentration gains even more relevance when the extremely crowded nature of the cellular interior is taken into account. The extent of crowding within the cell is exemplified by figures such as the estimated concentration of biomacromolecules in the *E. coli* cytosol ranging between 300-400 mg·ml^{-1} (Zimmerman & Trach 1991), or the volume of macromolecules being in the interval of 20-30% of the total volume of the cell interior (Ellis 2001). Macromolecular crowding is considered to increase dramatically the effective local protein concentration, as well as limiting the diffusibility of biomacromolecules. Accordingly, high molecular weight polymeric compounds, such as ficoll or polyethylene glycol, which allow to mimic the crowded intracellular environment have been show to accelerate aggregation rates (van den Berg et al. 1999; Munishkina et al. 2004).

The observed correlation between gene expression, in terms of experimentally determined mRNA levels, and decreasing predicted aggregation propensity in the human proteome (Tartaglia & Vendruscolo 2009) confirms protein abundance is tightly regulated by the cell in order to minimize the possibility of aberrant aggregation. However it is know that mRNA levels are not sufficient to justify the steady-state cellular concentration of polypeptides since post-transcriptional and co-translational processes play an important role in determining protein abundances (de Sousa Abreu et al. 2009). We have exploited several experimentally-derived datasets of mRNA levels, and protein abundance and solubility available for *E. coli*, in order to evaluate the previously observed relationship between mRNA levels and protein aggregation propensity, to assess whether such a trend extrapolates to a model organism such as *E. coli*, and to gain further insights regarding the cellular regulation of protein abundance as a function of the experimentally determined tendency to aggregate.

The analysis of mRNA levels obtained by different experimental methods shows that lower gene expression is associated to more aggregation-prone polypeptides, while highly expressed genes are linked to an increased protein solubility. In this sense, the solubility distributions of the proteins associated to the 25% lowest and 25% highest expressed genes are significantly different. These observations indicate that the control of polypeptide abundance at the transcription level, in the context of its association to the aggregative properties of proteins, which was observed for the human case is already present in bacteria, thus reinforcing the idea that the regulation of protein concentration is employed by the cell as an strategy to control the risk of aggregation. Nonetheless, as noted before, post-transcriptional and co-translational regulation implies that mRNA levels may not faithfully represent the real protein abundance in the cell. Several approaches have allowed to directly quantify protein abundance within the *E. coli* proteome (Lu et al. 2007; Ishihama et al. 2008), so the available data has allowed to analyze the aggregation properties of bacterial proteins as a function of their real abundances. Following the same trend observed for the mRNA levels, lower polypeptide abundance is related to an increased aggregation propensity; meanwhile the most abundant proteins exhibit a higher solubility. Again, the solubility distributions of the proteins with the 25% lesser and the 25% greater abundances show a clear difference, although slightly more significant that when mRNA levels are considered.

A more detailed analysis of polypeptide solubility reveals that the solubility distributions for proteins with high mRNA levels or abundance are almost coincident, while the distributions for polypeptides with low levels of mRNA or abundance are quite divergent. The original study of protein solubility lead by Taguchi and coworkers (Niwa et al. 2009) already noted a bimodal distribution for the aggregation propensity within the E. coli proteome. By taking into account the levels of gene expression and protein abundance, we find that such a bimodal distribution arises mostly from the solubility properties of proteins transcribed and translated at lower levels, since high mRNA and abundance levels correspond predominantly with the most soluble polypeptides. The coincidence in their solubility distributions suggests the concentration of highly abundant proteins is tightly regulated at the gene expression level, as the significant correlation between their cellular abundance mRNA levels confirms.

Altogether our analysis indicates that proteins the cell requires in high concentrations to develop their function are under higher selective pressure for more soluble sequences. In this case, their biosynthesis appears optimized to yield high concentrations through a precise regulation at the gene expression level. For low abundant proteins, their observed bimodal solubility distribution allows to postulate two different populations. In first place, aggregation-prone proteins the cell needs to maintain at low levels in order to avoid their aberrant deposition. On the other hand, the fraction of highly soluble but low abundant polypeptides might correspond to proteins whose functional concentration can experience significant increases under certain cellular conditions, thus they are also under selection to minimize their aggregation propensity. The poor correlation between their abundance and their mRNA levels indicates the cellular concentrations of low abundant proteins are controlled at different levels besides gene expression. In the case of aggregation-prone polypeptides, post-translational levels of regulation, and particularly the PQC and the degradation machineries, likely play a relevant role in defining their cellular abundance.

Overall, the observed trend provides additional support to the notion of protein aggregation propensity being a strong evolutionary constraint that shapes protein sequences, particularly for those required at high copy numbers. At the same time the cell has developed a series of mechanisms that allow the tight regulation of optimal protein abundances, so as to overcome deleterious aggregation.

DOI 10.1002/biot.201100014

Research Article

The aggregation properties of *Escherichia coli* proteins associated with their cellular abundance

Virginia Castillo Ricardo Graña-Montes* and Salvador Ventura*

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

Proteins are key players in most cellular processes. Therefore, their abundances are thought to be tightly regulated at the gene-expression level. Recent studies indicate, however, that steady-state cellular-protein concentrations correlate better across species than the levels of the corresponding mRNAs; this supports the existence of selective forces to maintain precise cellular-protein concentrations and homeostasis, even if gene-expression levels diverge. One of these forces might be the avoidance of protein aggregation because, in the cell, the folding of proteins into functional conformations might be in competition with anomalous aggregation into non-functional and usually toxic structures in a concentration-dependent manner. The data in the present work provide support for this hypothesis because, in *E. coli*, the experimental solubility of proteins correlates better with the cellular abundance than with the gene-expression levels. We found that the divergence between protein and mRNAs levels is low for high-abundance proteins. This suggests that because abundant proteins are at higher risk of aggregation, cellular concentrations need to be stringently regulated by gene expression.

Received 21 January 2011
Revised 22 February 2011
Accepted 24 February 2011

Supporting information available online

Keywords: *Escherichia coli* · Protein aggregation · Protein expression · Protein evolution · Protein folding

1 Introduction

Protein misfolding and aggregation are becoming central issues in biology and medicine. This is mainly because the failure of polypeptides to remain soluble in human cells is linked to the onset of more than 40 different disorders, ranging from Alzheimer's disease to diabetes [1–3]. The large number of usually redundant quality-control mechanisms and the considerable amount of energy the cell employs to assist proper protein folding reflects how important it is to prevent misfolding and aggregation in the cellular environment [4, 5].

It has been shown that the *in vitro* aggregation rates of several human proteins are not comparable with their gene-expression levels in vivo, as estimated from measurements of the cellular mRNA concentrations [6]. Computational analysis of the predicted aggregation propensities of large sets of human [7] and bacterial [8, 9] proteins from their primary sequences indicates that these theoretical values are also inversely correlated with the experimentally determined cellular concentrations of the corresponding mRNAs, thus suggesting that gene-expression levels are somehow linked to the solubility of the encoded protein or vice versa.

Although a strong correlation between gene-expression levels and cellular-protein abundance is usually assumed, downstream processing of mRNA and proteins introduces deviations in this relationship [10]. Only 20–60% of the differences in the steady-state concentrations of polypeptides are at-

Correspondence: Dr. Salvador Ventura
Institut de Biotecnologia i Biomedicina, Universitat Autònoma de Barcelona, 08193-Bellaterra (Barcelona), Spain
E-mail: salvador.ventura@uab.es

Abbreviations: APEX, absolute protein expression; **emPAI**, exponentially modified protein abundance index; **PSORT**, prediction of protein-sorting signals and localization sites in amino acid sequences

*These authors contributed equally to this work.

tributable to gene-expression levels, depending on the particular organism and conditions considered [11]. In *E. coli,* only 47% of the variance in protein abundance is explained by mRNA abundance. This rather weak correlation is likely to result from the bacterial operon structure, with genes obligatorily co-transcribed, but often differentially translated [12]. Marcotte and co-workers have recently analyzed large-scale protein and mRNA expression datasets of different species, from bacteria to human, and shown that the steady-state abundance of proteins correlates better across different phylogenetic taxa than the levels of the corresponding mRNAs [13]. This higher preservation of protein abundances might imply that the cellular concentrations of polypeptides are, to some extent, optimized and suggests the presence of selective pressures acting at the protein level to maintain optimal concentrations, regardless of divergent mRNA levels [13]. The maintenance of protein solubility levels that permit biochemical reactions in the cell, while avoiding deleterious aggregation, might be one of these forces. Protein aggregation is a multimolecular process that usually follows second-order kinetics [14]. In vitro aggregation is highly dependent on protein concentration and it is likely that the same principle would apply inside the cell. Therefore, the observed relationship between gene-expression levels and predicted protein-aggregation propensities might only reflect the existence ofextensive regulation, at the protein level, of cellular polypeptide abundances according to their particular solubility. To confirm this hypothesis, herein we exploit different *E. coli* databases to analyze the extent to which steady-state mRNA and protein levels correlate with experimental solubility.

2 Materials and methods

Taguchi and co-workers could quantify experimentally the amount of in vitro translated protein in the soluble and insoluble fractions of 3173 different *E. coli* proteins [15]. From these results, we have selected a subset of 597 different polypeptides for which the cellular-protein abundances have been experimentally measured. Beginning with the Taguchi database of solubility, we constructed two different protein subsets for subsequent analysis. The first database (Supporting information, Table SL) includes 495 cytosolic *E. coli* proteins that fulfil the following criteria: (i) they have been experimentally located in the cytoplasm [16] and theoretically analyzed by using the PSORT (prediction of protein-sorting signals and localization sites in amino acid sequences) algorithm [17], available at http://db.psort.org/, to be considered cytosolic; (ii) their cellular concentrations have been experimentally determined by mass spectrometry using the emPAI (exponentially modified protein abundance index) approach [12, 16]; and (iii) the relative levels of the mRNAs encoding most of these proteins have been quantified by using two-color fluorescent DNA microarrays [18]. The second database (Table S2 in the Supporting information) includes 335 *E. coli* proteins residing in different subcellular compartments for which the cellular concentrations have been experimentally determined by mass spectrometry using the APEX (absolute protein expression) methodology [12] and for which mRNA levels have been reported in the same study [12]. There are 233 proteins shared by both datasets.

‘High-abundant’ molecules are considered here as the 25% of proteins with the highest associated mRNA or protein levels. Accordingly, the 25% of proteins with the lowest associated mRNA or protein levels are considered here as ‘low-abundant’ molecules. Proteins for which <30% of the polypeptide is present in the soluble fraction when they are translated in vitro, as determined by Taguchi and co-workers [15], are considered to be ‘aggregation-prone’ or ‘insoluble’. Proteins for which >70 % of the polypeptide is present in the soluble fraction when they are translated in vitro are considered to be ‘soluble’.

The protein datasets were sorted according to mRNA or protein levels and the solubility properties of high- and low-abundant molecules according to these criteria were compared. Alternatively, the datasets were sorted according to experimental protein solubility and the mRNAs and protein abundances of insoluble and soluble polypeptides were compared.

To obtain the averaged distribution of protein solubility according to their mRNA and protein levels (shown in Fig. 3A), the proteins in the two databases were ordered according to either their mRNA or protein levels. Then, the 100 proteins with the highest associated protein or mRNA levels in each database were joined to form the averaged high-abundant mRNA and protein subsets. The same procedure was followed to obtain the low-abundant averaged mRNA and protein subsets.

The solubility distribution of the different protein subsets (shown in Fig. 3B and 3C) was analyzed by fitting the data to a normal or a double-Gaussian function by using a non-linear least-squares curve fitting with the KaleidaGraph program (Synergy Software). This software was also used for the rest of the statistical analysis.

To analyze the correlation between experimental mRNA or protein concentrations and protein solubility (shown in Fig. 6 below), the complete datasets were binned into groups, calculated by rank-ordering mRNAs and proteins by protein solubility and calculating average solubility, protein and mRNA expression levels per bin of five proteins.

3 Results and discussion

3.1 Databases

Protein aggregation in bacteria has long been regarded as an unspecific process that relies on the establishment of hydrophobic contacts between partially folded intermediates. However, recent studies have converged to demonstrate unequivocally that bacterial aggregates share a number of common features with the pathogenic amyloid fibrils linked to human diseases, including cytotoxicity and infectivity in the case of prion proteins [19–23]. Therefore, bacteria, and specifically *E. coli,* constitute a simple but physiologically relevant system in which to study and model protein aggregation. In this context, Taguchi and co-workers have individually synthesized almost the entire ensemble of *E. coli* proteins by using an in vitro reconstituted translation system and experimentally characterized the individual aggregation properties [15], thus providing an experimental benchmark with which to analyze the relationship between solubility and other protein properties.

Because the aggregation propensities of proteins appear to depend on the subcellular compartment in which they reside [24], in the present study we considered two different protein databases. The first one corresponds to 495 cytosolic proteins in the Taguchi database [15], for which Ishihama and co-workers have succeeded in determining the experimental abundance using the emPAI approach [16, 25]. The second one does not consider protein localization and corresponds to 335 polypeptides in the Taguchi database, for which Marcotte and co-workers have measured the absolute protein levels in *E. coli* using the APEX methodology [12]. We also expected that the use of protein-abundance datasets calculated by two different experimental methods would minimize methodological bias in the analysis. These protein-abundance data, together with the previous experimental quantification of the gene-expression levels in *E. coli* cells under equivalent conditions [12, 18], provide a new benchmark with which to analyze how the bacterial transcriptome and proteome composition, and more specifically mRNA and protein abundances, are related to protein solubility. Overall, for the present study we analyzed around 600 different bacterial proteins, for which experimental solubility, associated mRNA levels, and protein concentrations were available in the above-mentioned databases (see the Materials and Methods section and the Supporting information).

3.2 Relationship between mRNA levels and protein solubility

We first analyzed the solubility of proteins with the highest and lowest associated mRNA levels according to the data obtained by Cohen and co-workers by using two-color fluorescent DNA microarrays [18]. A comparison of the solubility distribution of proteins in these subsets is shown in Fig. 1, which illustrates that, as a trend, the proteins encoded by high- and low-expressed genes exhibit differential aggregation properties (p<0.0002 unpaired Wilcoxon). High-expressed mRNAs correspond overall to more soluble proteins than mRNAs present at low levels in the cell. Marcotte and co-workers have shown that averaging the absolute mRNA levels obtained by using different technologies tends to remove technology-specific errors and renders more consistent data [12]. We therefore analyzed the solubility distribution of proteins displaying high and low expression levels according to the average mRNA levels measured by at least two of the three different methods [12]. The cumulative frequency curves obtained with this averaged database do not differ significantly from that obtained with only DNA microarray data (Fig. 1); this indicated again that proteins with associated high- and low-expressed mRNA levels exhibit different experimental aggregation properties (p<0.004 unpaired Wilcoxon). Overall, the data are in good agreement with previous results obtained by comparing experimental mRNA levels and theoretical protein-aggregation propensities, as calculated computationally from the primary sequences [7–9].

3.3 Relationship between protein levels and protein solubility

There is increasing evidence that due to post-transcriptional, translational, and stability regulation, mRNA abundance alone does not necessarily reflect the steady-state protein abundance in a cell [11, 26]. Therefore, the observed relationship between mRNA abundance and protein solubility should not automatically be assumed to occur at the proteome level.

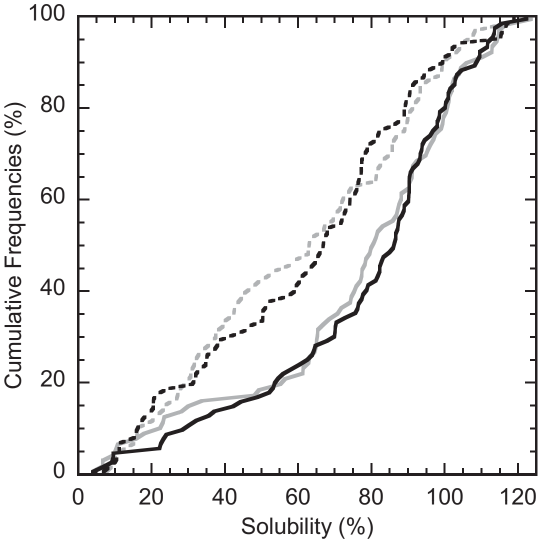


Figure 1. The relationship between mRNA levels and protein solubility. A cumulative distribution of solubility for mRNA levels associated with cytosolic and total *E. coli* proteins; mRNA levels were retrieved from the groups of Cohen [18] (black) and Marcotte [12] (gray) databases, respectively. The solubility for the 25% less abundant mRNAs are shown as dashed lines, and for the 25% more abundant mRNAs as solid lines.

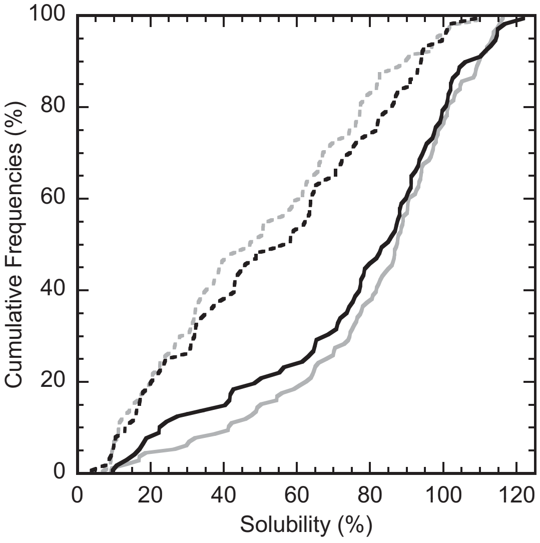


Figure 2. The relationship between protein levels and protein solubility. A cumulative distribution of solubility according to cytosolic and total protein levels, as derived from emPAI [16] (black) and APEX [12] (gray) analysis, respectively. The solubility for the 25% less expressed proteins are shown as dashed lines, and for the 25% more expressed proteins as solid lines.

Figure 2 shows the solubility distribution of proteins found at high and low copy numbers in *E. coli*, according to both emPAI and APEX analyses. Independent of the approach, the cumulative frequency curves obtained for the high- and low-abundant protein subsets are similar, with high-abundant proteins displaying higher solubility than low-abundant ones. Therefore, at the protein level, the solubility distribution of proteins differing in their steady-state abundance is also significantly different (p<0.0001 unpaired Wilcoxon), independent of the quantification method and if we consider only cytosolic proteins or do not take into account protein location.

3.4 Abundant mRNAs and proteins exhibit similar solubility distributions

To compare if the solubility properties of the high- and low-abundant polypeptide sets resemble that of proteins with high and low gene-expression levels, we averaged the results obtained for mRNAs and proteins in the two databases and analyzed the solubility properties of the molecules in the resulting high- and low-abundant groups (Fig. 3A). The frequency curves for the solubility of high-expressed mRNAs and high-abundant proteins are very similar, with mean solubilities of 75 and 78 %, respectively. In contrast, the solubility properties of low-expressed mRNAs and low-abundant proteins are clearly differentiable, displaying mean solubilities of 51 and 61 %, respectively. These differences are more easily visualized when the solubility distributions in the two protein groups are represented as histograms (Fig. 3B). Taguchi and co-workers demonstrated that the solubility of the *E. coli* proteome is not homogenously distributed across a continuum of values corresponding to a normal distribution, but rather that most polypeptides can be classified into one of two groups: those with a high aggregation propensity and those that are highly soluble, corresponding thus to a bimodal distribution [15]. This property is propagated in the solubility distribution of low-expressed mRNAs and low-abundant proteins (Fig. 3B). However, in spite of the fact that aggregation-prone and soluble proteins are well represented in the two subsets, for each considered solubility range, the percentages of low-abundant proteins and low-expressed mRNAs are different (Fig. 3B). In contrast, the solubility distributions of high-expressed mRNAs and high-abundant proteins are coincident, since both are exceedingly enriched in soluble proteins, with only a minor population of aggregation-prone polypeptides (Fig. 3C). Therefore, whereas low-abundant proteins can display a broad range of solubilities, high-abundant proteins are preferentially soluble. This observation supports the theory that the prevention of protein aggregation in biological environments — a phenomenon that is strongly dependent on the local polypeptide concentration — may act as an important evolutionary

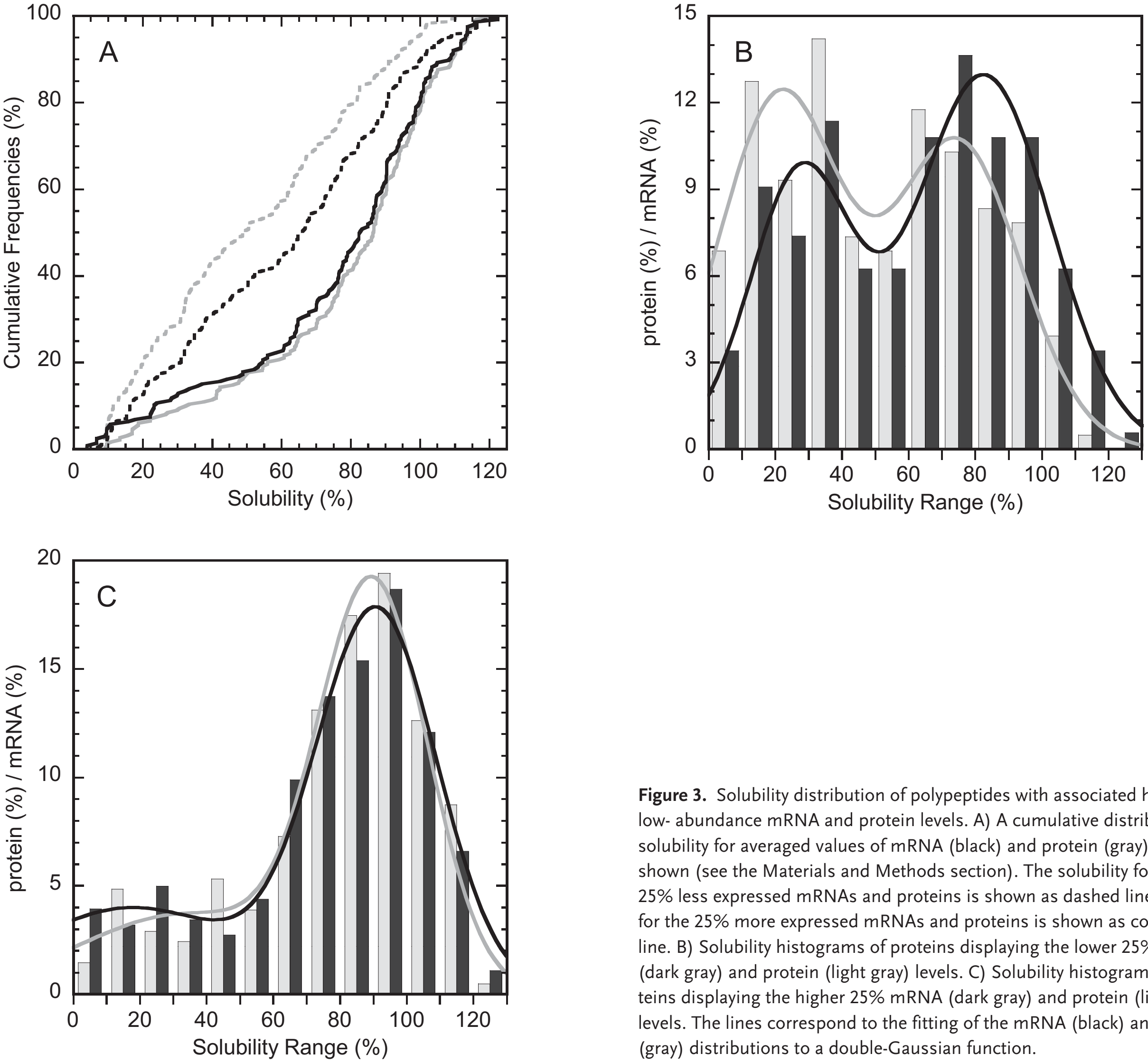


Figure 3. Solubility distribution of polypeptides with associated high- and low- abundance mRNA and protein levels. A) A cumulative distribution of solubility for averaged values of mRNA (black) and protein (gray) levels is shown (see the Materials and Methods section). The solubility for the 25% less expressed mRNAs and proteins is shown as dashed lines, and for the 25% more expressed mRNAs and proteins is shown as continuous line. B) Solubility histograms of proteins displaying the lower 25% mRNA (dark gray) and protein (light gray) levels. C) Solubility histograms of proteins displaying the higher 25% mRNA (dark gray) and protein (light gray) levels. The lines correspond to the fitting of the mRNA (black) and protein (gray) distributions to a double-Gaussian function.

constraint during protein evolution. If the maximum cellular concentration of a protein depends on its solubility, highly expressed proteins should have evolved particularly low aggregation propensities to remain soluble. Therefore, it is not surprising that over-expression of proteins above their natural concentrations by using homologous and especially heterologous recombinant systems produce, in many cases, the intracellular accumulation of the target polypeptides as aggregated species [27], which is likely to be the result of exceeding naturally evolved solubility limits.

3.5 Abundant proteins exhibit low divergence between protein and mRNA levels

A feasible explanation for the low divergence in solubility between high-expressed genes and high-abundant proteins and the significant difference in the solubility distribution of proteins with associated low mRNA levels and those exhibiting low protein concentrations, would be that, for high-abundant proteins, the steady-state polypeptide concentrations and the absolute mRNA levels would correlate much better than those of low-abundant molecules. Figure 4 shows that, independent of the selected databases, this is the case, with a highly significant correlation for high-abundant proteins (p<0.0000001 paired Wilcoxon) and a non-significant correlation for low-abundant ones (p<0.4 paired Wilcoxon, in the best case). Therefore, even though the present dataset is small and represents only about 18 % of the total *E. coli* proteome, the data suggest that the previously observed divergence between protein and mRNA levels [13] is not uniform for all of the proteins in an organism, with

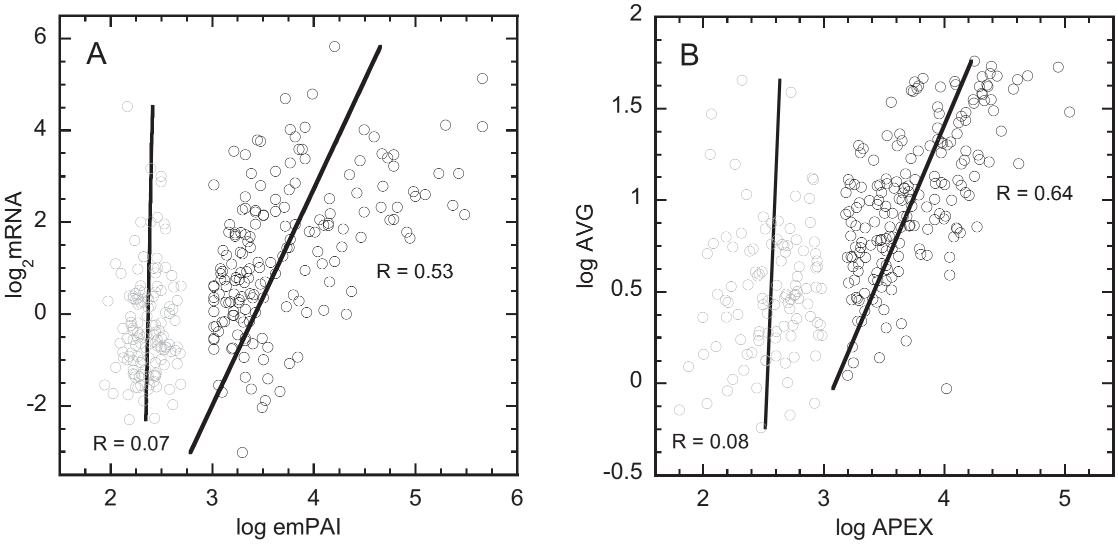


Figure 4. Correlation between mRNA and protein concentrations according to the number of cellular-protein copies. A) Cytosolic protein levels as quantified by emPAI [16] and mRNA levels, as reported by Cohen and co-workers [18]. B) Total protein levels as quantified by APEX [12] and mRNA levels, as reported by Marcotte and co-workers [12]. Data of proteins displaying less than 1000 copies per cell are shown in gray. Data of proteins displaying more than 1500 copies per cell are shown in black.

the abundance of some proteins being tightly regulated at the gene-expression level and the concentration of other polypeptides displaying high divergence relative to their mRNA levels. Interestingly enough, when the transcriptomes and proteomes of fly and nematode were compared, a contribution of the expression level to the correlation between mRNA and protein levels was also observed. High-expressed proteins and mRNAs were found to be similarly conserved in their concentrations across the two species, whereas the divergence in the coefficient correlation was high for the least-abundant proteins and mRNAs [13, 28]. Our analysis suggests that protein-aggregation propensity might be an important factor contributing to this differential regulation because, at least in our protein collection, the correlation between mRNA and protein levels is higher for highly soluble proteins than for the rest of proteins, independent of the considered database and approach (Fig. 5).

3.6 Cellular-protein levels correlate better with protein solubility than gene-expression levels.

Protein activities and concentrations, and not mRNA levels, are the real factors responsible for cellular physiology at any time and any given cell conditions. Therefore, one should expect that if protein solubility acts as an evolutionary constraint, its effects would become more evident in protein abundance than in gene-expression levels. To explore this possibility, we analyzed the relationship between abundance and experimental solubility for mRNA and protein levels in our dataset (Fig. 6). A highly significant correlation was observed (p<0.0000001 paired Wilcoxon) with R=0.68 and R=0.57 between protein solubility and the number of polypeptide copies per cell, according to the emPAI and APEX databases, respectively. The higher correlation in the emPAI database might result from the fact that all of the proteins in this group correspond to cytosolic proteins, whereas those in the APEX database have a diverse origin, since the cellular localization of a given protein has been shown to be a factor influencing its solubility properties [9, 29, 30].

Although the correlation between mRNA levels and solubility was also statistically significant, it turned out to be weaker (p<0.0001 paired Wilcoxon, in the best case) with R values of 0.43 and 0.38 for the averaged and microarray mRNA abundance measurements, respectively.

4 Concluding remarks

Overall, the present analysis suggests that the real cellular-protein concentrations, rather than gene-expression levels, co-evolve with protein solubility. This relationship is likely to result from evolutionary constraints to decrease the aggregation of natural proteins at the maximum concentrations required to accurately perform their functions [6], maintaining protein homeostasis, and cellular fit-

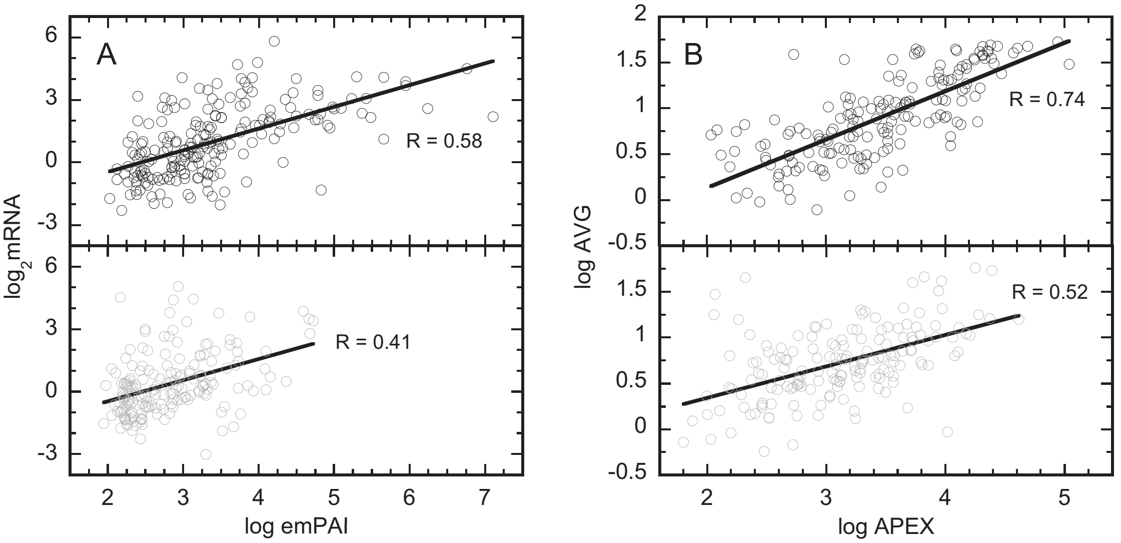


Figure 5. Correlation between mRNA levels and protein abundance according to experimental protein solubility. A) Cytosolic protein levels as quantified by emPAI [16] and mRNA levels, as reported by Cohen and co-workers [18]. B) Total protein levels as quantified by APEX [12] and mRNA levels, as reported by Marcotte and co-workers [12]. Data of proteins displaying solubility lower than 70 %, as measured by Taguchi and co-workers [15], are shown in gray. Data of proteins displaying solubility higher than 70 % are shown in black.

ness. In this sense, highly abundant proteins tend to be highly soluble and their abundance seems to be tightly regulated at the gene-expression level, which might reflect that the whole process of protein synthesis is optimized to produce high amounts of these proteins. In this way, theoretical models indicate that, to maintain protein homeostasis in the cell, the ratio of protein-aggregation propensity should scale proportionally with the cellular-protein concentration and, as a consequence, that higher protein abundance implies higher evolutionary pressure [31]. Because the aggregation propensity of proteins depend both on their sequence and fold — two features well preserved between homologous proteins in different species [32] — it is not surprising that the steady-state abundances of proteins correlate across different phylogenetic taxa, especially for high-abun-

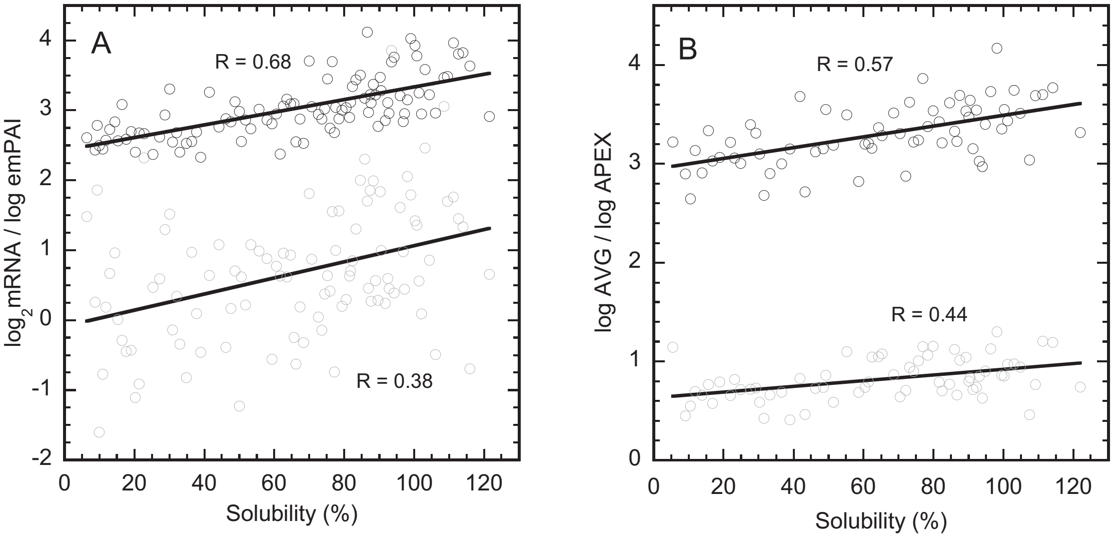


Figure 6. Correlation of mRNA and protein levels with experimental protein solubility. The complete cytosolic (A) and total (B) protein datasets were binned after rank-ordering proteins by solubility and calculating average solubility, mRNA and protein levels per bin of five proteins. The correlation between mRNA levels and protein solubility is shown by using gray symbols. The correlation between protein levels and protein solubility is shown by using black symbols.

dant polypeptides [13]. For low-abundant proteins, post-translational regulation of protein concentrations must compensate for the divergent mRNA expression levels to maintain these polypeptides at functional levels, especially for essential proteins, such as cell-cycle regulators, transcription factors, or proteins involved in chromosome segregation, which are known to be present at very low levels. Also, proteins usually found at low concentrations, but for which abundance can increase dramatically under certain cellular conditions, are likely to have evolved optimized solubility properties and might constitute a fraction of the low-abundant but highly soluble proteins detected in our dataset. Overall, the present study provides additional evidence to support the hypothesis that the aggregation properties of polypeptides have been shaped by evolution according to specific cellular requirements.

We acknowledge Prof. Hideki Taguchi and his group for sending us the crude expression and solubility data from their in vitro reconstituted translation system expression experiment. This study was supported by a grant BFU2010-14901 from Ministerio de Ciencia e Innovación (MCI-Spain) and grant 2009-SGR 760 from AGAUR (Generalitat de Catalunya). V.C. and R.G.M. are beneficiaries of predoctoral FPU AP2007-02849 and FPU AP2009-0948 fellowships from the Ministerio de Educación-Spain, respectively. S.V. has been granted an ICREA ACADEMIA award (ICREA).

The authors have declared no conflict of interest.

5 References

[1] Luheshi, L. M., Dobson, C. M., Bridging the gap: from protein misfolding to protein misfolding diseases. *FEBS Lett.* 2009, *583*, 2581–2586.

[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] Cohen, E., Bieschke, J., Perciavalle, R. M., Kelly, J. W., Dillin, A., Opposing activities protect against age-onset proteotoxicity. *Science* 2006, *313*, 1604–1610.

[4] Bukau, B., Weissman, J., Horwich, A., Molecular chaperones and protein quality control. *Cell* 2006, *125*, 443–451.

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

[6] 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.

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

[8] 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.

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

[10] Anderson, L., Seilhamer, J., A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* 1997, *18*, 533–537.

[11] de Sousa Abreu, R., Penalva, L. O., Marcotte, E. M., Vogel, C., Global signatures of protein and mRNA expression levels. *Mol. BioSyst.* 2009, *5*, 1512–1526.

[12] Lu, P., Vogel, C., Wang, R., Yao, X., Marcotte, E. M., Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* 2007, *25*, 117–124.

[13] Laurent, J. M., Vogel, C., Kwon, T., Craig, S. A., et al., Protein abundances are more conserved than mRNA abundances across diverse taxa. *Proteomics* 2010, *10*, 4209–4212.

[14] 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.

[15] Niwa, T., Ying, B. W., Saito, K., Jin, W., et al., Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of *Escherichia coli* proteins. *Proc. Natl. Acad. Sci. USA* 2009, *106*, 4201–4206.

[16] Ishihama, Y., Schmidt, T., Rappsilber, J., Mann, M., et al., Protein abundance profiling of the Escherichia coli cytosol. *BMC Genomics* 2008, *9*, 102.

[17] Gardy, J. L., Laird, M. R., Chen, F., Rey, S., et al., PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics* 2005, *21*, 617–623.

[18] Bernstein, J. A., Khodursky, A. B., Lin, P. H., Lin-Chao, S., Cohen, S. N., Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *Proc. Natl. Acad. Sci. USA* 2002, *99*, 9697–9702.

[19] Doglia, S. M., Ami, D., Natalello, A., Gatti-Lafranconi, P., Lotti, M., Fourier transform infrared spectroscopy analysis of the conformational quality of recombinant proteins within inclusion bodies. *Biotechnol. J.* 2008, *3*, 193–201.

[20] 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.* 2009, *48*, 4858–4860.

[21] 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.

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

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

[24] Vendruscolo, M., Tartaglia, G. G., Towards quantitative predictions in cell biology using chemical properties of proteins. *Mol. BioSyst.* 2008, *4*, 1170–1175.

[25] Ishihama, Y., Oda, Y., Tabata, T., Sato, T., et al., Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of

sequenced peptides per protein. *Mol. Cell. Proteomics* 2005, *4*, 1265–1272.

[26] Kudla, G., Murray, A. W., Tollervey, D., Plotkin, J. B., Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* 2009, *324*, 255–258.

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

[28] Schrimpf, S. P., Weiss, M., Reiter, L., Ahrens, C. H., et al., Comparative functional analysis of the Caenorhabditis elegans and Drosophila melanogaster proteomes. *PLoS Biol.* 2009, *7*, e48.

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

[30] Monsellier, E., Ramazzotti, M., Taddei, N., Chiti, F., Aggregation propensity of the human proteome. *PLoS Comput. Biol.* 2008, *4*, e1000199.

[31] Pechmann, S., Vendruscolo, M., Derivation of a solubility condition for proteins from an analysis of the competition between folding and aggregation. *Mol. BioSyst.* 2010, *6*, 2490–2497.

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

CHAPTER 5.- FUNCTIONAL CONSTRAINTS LIMITING SELECTION AGAINST PROTEIN AGGREGATION

5.1.- Introduction

Despite the evolutionary selection against aggregation is now a well recognized phenomenon (Tartaglia et al. 2005), the persistent aggregative load observed for a majority of protein sequences (Reumers et al. 2009) suggests that such purifying selection to attain more soluble sequences is limited to a certain extent, likely because the maintenance of protein functionality requires of sequential properties that overlap with those that define APRs.

One of such functional constraints might be the capability of a protein sequence to fold into a defined three-dimensional conformation. Because the forces driving both the folding into a functional native state and the aggregation into aberrant conformations appear, at least during the initial stages of these processes, to be of a similar nature, folding and aggregation are usually considered competing reactions. The frequent burial of APRs within the native structure of globular proteins (Linding et al. 2004; Buck et al. 2013) is, as well, taken as an indication of the competition between folding and aggregation. The competition between these two processes suggests that the sequential determinants accounting for the emergence of the propensity to aggregate overlap, at least partially, with those guiding the folding into the native structure. Accordingly, selection against aggregation propensity might have a detrimental effect on the folding efficiency of a protein. Thus, foldability may restrain the evolutionary pressure to minimize aggregation propensity, also because the burial of APRs within the native structure is considered to possess a protective role against deleterious aggregation.

Additional constraints may also arise from the requirement to adopt a particular quaternary structure or to establish functional protein-protein interactions since, as detailed in the previous chapter and exemplified by the analysis of the SUMO domains, the properties of the amino acids involved in the formation of protein-protein interfaces resemble those of the residues promoting aggregation propensity.

Aside from the structural effects on the maintenance of protein function, selection against aggregation may also alter the environment of specific residues exerting a catalytic activity, the extent of this source of functional constraint for the evolutionary pressure to prevent aggregation is discussed below more in-depth.

In the following sections, the aggregative properties of two homogeneous ensembles of proteins are described, allowing to dissect specific functional constraints from other factors inducing dissimilar selective pressure against aggregation propensity across proteomes.

5.2.- Analysis of Specific Aggregation Properties of Functional Classes from Different Organisms. Insights from the Kinase Complements of Proteomes

The evolutionary pressure against protein aggregation does not act uniformly over all the proteins of a given proteome. As discussed in the previous chapter, variables such as the functional cellular abundance or the *in vivo* lifetime of a protein influence the strength of selection for more soluble polypeptide sequences. It has also been observed that the functional role of a protein also influences the evolutionary pressure exerted over it, with proteins exerting essential functions being under higher selection against aggregation (Chen & Dokholyan 2008). In this sense, the analysis of the aggregation properties of a homogeneous functional class of proteins, and its comparison between different organisms, allows to dissect the specific effects of the selective pressure against protein aggregation in this class from other variables acting generically on proteomes. At the same time, focusing the analysis on an ensemble of proteins expected to be under similar selective pressure may help to identify those constraints imposing limitations to the purification of aggregation-prone sequences during evolution.

Protein kinases are an extremely relevant class of proteins, which are involved in virtually every regulatory process within the cell. In addition, the kinase domains in the majority of protein kinase groups share structural similarity. These two properties make this class of proteins an outstanding ensemble in order to survey the impact of evolutionary selection against protein aggregation and to detect functional constraints limiting that purifying pressure. We have exploited a database of the kinase complement of proteomes, or kinomes, from different organisms where the boundaries of the specific kinase domains within the entire kinase proteins have been determined. The average aggregation propensities of the ensembles of kinase domains from different species show a decreasing trend as organism complexity rises, which is comparable to the evolutionary distances between those species. This finding confirms the same tendency previously observed employing whole proteomes (Tartaglia et al. 2005). Interestingly, in the case of kinase domains, selection against aggregation does not arise merely from the global reduction of the number of APRs detected in their sequences but rather from a decrease in the overall aggregation propensity resulting from a balance between the number of “hot-spots” and its aggregation potential.

A detailed analysis of the relationship between the aggregation properties of the human kinase domains and the entire proteins they belong to, reveals that the domains possess a higher tendency to aggregate than complete proteins. The evaluation of the aggregation along entire kinase proteins explains why kinase domains present a higher propensity, since APRs are more concentrated and exhibit a greater aggregation potential within domains than in the rest of the molecule. This suggests kinase domains face higher functional constraints restricting the selection against aggregation than the remainder of the proteins harboring them. Such functional constraints are further evidenced when analyzing the aggregation propensities of kinase functional groups separately; for example, the STE and CAMK groups exhibit highly dissimilar aggregation properties. While STE kinases function transducing signals from the cell surface to the nucleus, thus being under a specific pressure for increased solubility, kinases from the CAMK group usually associate to form complexes, so they confront particular restrains limiting the evolutionary pressure to minimize aggregation, because the formation of functional oligomerization interfaces is driven by similar forces than those promoting aggregation (Pechmann et al. 2009; Castillo & Ventura 2009). However, human protein kinases from all groups present very low aggregation propensities, and no evident trend is observed for the correspondence of aggregation properties between domains and entire protein within each kinase group. It appears therefore that, independently of the functionally-constrained variability in the aggregation propensities presented by each particular class of kinase domain, protein kinases have generally evolved a reduced tendency to aggregate in order to compensate for the intrinsic aggregational load of the kinase domains.

In a global perspective, the inspection of the regions where APRs map over kinase domains allows to identify two primary sources of functional constraint. First, the majority of the amyloidogenic stretches detected in human kinase domains overlap with regular secondary structure elements of the canonical domain structure. In second place, in the kinase domains from different organisms, a significant number of the conserved catalytic residues which possess a key role for kinase function are embedded in or in close proximity to APRs. These observations indicate that the properties of the amino acids contributing to the aggregation propensity of polypeptides are also important for the attainment of a stable three-dimensional structure or to provide an appropriate environment for certain proteins to develop their catalytic functions. In the later case, this is further supported by a global analysis of the environments of catalytic residues (Buck et al. 2013), although catalytic amino acids do not tend to reside within aggregation-prone segments since they usually present a polar or charged character, these residues are more commonly found in close proximity with APRs that it would be expected by random chance. These functional requirements necessarily restrain the evolutionary selection to decrease the aggregation propensity of polypeptides.

An interesting possibility to explain the high aggregative load of kinase domains would be that these domains may have exploited chaperone binding as a functional constraint to improve their folding efficiently. It has been shown that although kinase domains can evolve away from chaperone-aided folding, their sequences have not generally followed this trend, as an example, about 60% of human kinases are assisted by the HSP90 chaperone (Taipale et al. 2010). Consequently, the maintenance of a certain aggregation propensity through chaperone binding determinants could serve as a functional strategy kinases have developed in order to ensure fidelity in the regulatory pathways they are involved in.

These functional constraints acting on kinase domains provide them with an inherent aggregational load that increases their risk of aberrant deposition. Since whole kinase proteins possess a significantly lower tendency to aggregate than their corresponding kinase domains, it appears that the inclusion of these domains into a greater proteinaceous entity could have served as a strategy to “buffer” their implicit aggregation propensity. The lower aggregation propensity in entire protein kinases correlates with a greater predicted intrinsic disorder, this indicates that the “buffering” effect of the aggregative potency of kinase domains is mediated by an enrichment in IDPRs in the rest of protein, whose impact on the tendency of polypeptides to aggregate has been addressed in the previous chapter.

frontiers in
PHYSIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 20 November 2012
doi: 10.3389/fphys.2012.00438

Protein aggregation profile of the human kinome

Ricardo Graña-Montes*[1], *Ricardo Sant'Anna de Oliveira*[2] and *Salvador Ventura*[1]

[1] *Institut de Biotecnologia i de Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain*
[2] *Instituto de Bioquímica Médica, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil*

Research Topic: ***From structural to molecular systems biology: experimental and computational approaches to unravel mechanisms of kinase activity regulation in cancer and neurodegeneration***

Edited by:
Matteo Barberis, Humboldt University Berlin, Germany; Max Planck Institute for Molecular Genetics, Berlin, Germany

Reviewed by:
David James, Garvan Institute of Medical Research, Australia
Paolo Tortora, University of Milano-Bicocca, Italy

****Correspondence:***
Salvador Ventura, Institut de Biotecnologia i de Biomedicina, Departament de Bioquímica i Biologia Molecular, Parc de Recerca UAB, Mòdul B, Universitat Autònoma de Barcelona, Bellaterra (Barcelona) E-08193, Spain.
e-mail: salvador.ventura@uab.es

Protein aggregation into amyloid fibrils is associated with the onset of an increasing number of human disorders, including Alzheimer's disease, diabetes, and some types of cancer. The ability to form toxic amyloids appears to be a property of most polypeptides. Accordingly, it has been proposed that reducing aggregation and its effect in cell fitness is a driving force in the evolution of proteins sequences. This control of protein solubility should be especially important for regulatory hubs in biological networks, like protein kinases. These enzymes are implicated in practically all processes in normal and abnormal cell physiology, and phosphorylation is one of the most frequent protein modifications used to control protein activity. Here, we use the AGGRESCAN algorithm to study the aggregation propensity of kinase sequences. We compared them with the rest of globular proteins to decipher whether they display differential aggregation properties. In addition, we compared the human kinase complement with the kinomes of other organisms to see if we can identify any evolutionary trend in the aggregational properties of this protein superfamily. Our analysis indicates that kinase domains display significant aggregation propensity, a property that decreases with increasing organism complexity.

Keywords: protein kinases, protein aggregation, amyloid, protein evolution, AGGRESCAN

INTRODUCTION

Most polypeptides need to fold into specific three-dimensional structures to perform their biological functions (Dobson, 2003). Only the correctly folded forms of these proteins remain soluble in the cell and are able to interact with their molecular targets (Daggett and Fersht, 2009). Therefore, protein misfolding impairs cell fitness and is being found linked to an increasing number of human degenerative diseases. In these disorders, misfolded conformers establish non-native intermolecular contacts that result in their deposition into insoluble amyloid aggregates in the intra- or extracellular space (Chiti and Dobson, 2006). All these assemblies display a common cross-β motif (Nelson and Eisenberg, 2006). However, the ability to form amyloid-like structures is not restricted to a subset of disease-linked proteins and this conformation may be accessed by most, if not all, proteins in living organisms, from bacteria to human, irrespective of their native fold (Dobson, 2004; Jahn and Radford, 2005; de Groot et al., 2009). In fact, the molecular interactions leading to the formation of amyloids are similar to those promoting the folding and functional assembly of proteins (Linding et al., 2004; Castillo and Ventura, 2009). As a result, folding and aggregation pathways are continuously competing in the cell.

Globular proteins are soluble in their biological environments. However, when their stability is compromised by genetic mutations or environmental conditions, local unfolding might promote the exposure to the solvent of aggregation-prone regions previously protected in the native state (Ventura et al., 2004; Ivanova et al., 2006). To confront this danger, protein sequences have evolved strategies to reduce aggregation propensity (Rousseau et al., 2006b; Monsellier and Chiti, 2007; Tartaglia et al., 2007; Castillo et al., 2011). The selective pressure against protein aggregation is stronger for proteins involved in essential cellular functions (Tartaglia and Caflisch, 2007; Chen and Dokholyan, 2008; de Groot and Ventura, 2010), as it might be the case of kinases. Nevertheless, even those globular proteins selected to be highly soluble cannot avoid the presence of aggregation "sensitive" stretches in their sequences (Sabate et al., 2012). Accordingly, we have recently shown that a cancer associated point mutant of human nucleoside diphosphate kinase A displaying reduced conformational stability forms amyloid fibrils under close to physiological conditions (Georgescauld et al., 2011). Similarly, a partially folded intermediate of phosphoglycerate kinase has been shown to self-assemble into amyloid fibrils (Damaschun et al., 2000; Agocs et al., 2010).

The aggregation propensities of proteins are determined to a large extent by their sequences and the intrinsic properties that govern the aggregation of proteins have been already identified. This has allowed the development of a set of algorithms able to predict aggregation-prone regions in protein sequences as well as the overall aggregation propensity of polypeptides (Castillo et al., 2010; Hamodrakas, 2011). Several of these programs are well suited for the analysis of large protein sets, among them AGGRESCAN, an algorithm previously developed by our group (Conchillo-Sole et al., 2007; de Groot et al., 2012), which displays a high power to predict *in vivo* protein aggregation (Belli et al., 2011). Different predictive algorithms have been used to analyze the overall aggregation properties of complete proteomes, from bacteria to human (Tartaglia et al., 2005; Rousseau et al., 2006b; Monsellier et al., 2008; de Groot and Ventura, 2010). Here we address the intrinsic aggregational properties of protein sequences belonging to the same super-family in different organisms. In this way, we have analyzed the aggregation propensities of the protein kinase complements ("kinomes") of budding yeast, fly, mouse, and humans using AGGRESCAN.

RESULTS

AGGREGATION PROPERTIES OF KINOMES

We computed the aggregation properties of the complete kinomes of *S. cerevisiae* (104 proteins, all containing a single kinase domain), *D. melanogaster* (197 domains in 194 proteins), *M. musculus* (520 domains in 511 proteins), and *H. Sapiens* (508 domains belonging to 497 proteins) using AGGRESCAN. The following parameters were calculated (**Figure 1** and "Materials and Methods"):

1. The average aggregation propensity of the sequence (Na4vSS).
2. The frequency of occurrence of aggregation-prone regions (APR), i.e., the number of aggregating peaks for each 100 protein residues (NnHS).
3. The average aggregating potency of the detected aggregation peaks (THSAr), i.e., the area of the peaks that lies above the detection threshold, normalized by the protein length.
4. The average aggregating potency of residues above the detection threshold (AATr), independently if they are clustered in aggregating peaks or not, i.e., the area of the surface above the detection threshold.

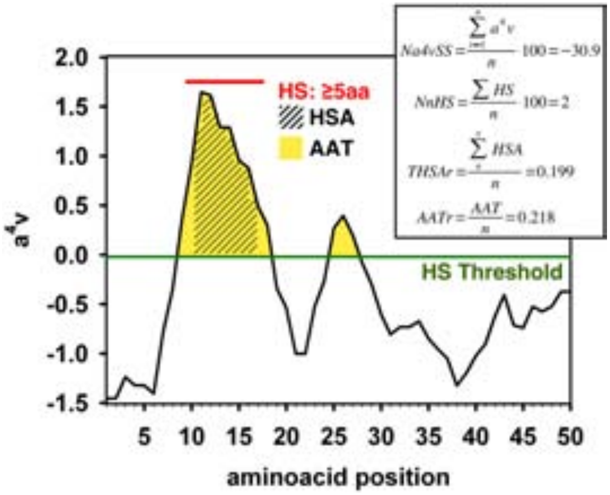


FIGURE 1 | Example of the AGGRESCAN output with the different parameters calculated by this algorithm. The aggregation profile is represented as the value of the experimentally derived parameter a^{4}v (de Groot et al., 2005) plotted against the query sequence. An overall value of this parameter is given for the whole sequence by dividing the a4v sum for all positions over the number of amino acids an multiplying by 100 (Na4vSS). An aggregation-prone region (APR) or Hot-Spot (HS) is detected whenever a stretch of 5 or more consecutive amino acids, none of them being a proline, with a4v values above the HS Threshold (HST) is found. The normalized number of Hot Spots (NnHS) is calculated as the number of Hot-Spots detected divided by the sequence length and multiplied by 100. The area of the profile above the threshold gives an idea of the aggregation potential of the sequence or a certain region; it can be calculated for the whole profile (AAT) or limiting it to the detected Hot-Spots (HSA). The area values are normalized dividing by the number of amino acids in the input sequence.

We retrieved the sequences corresponding to kinase domains in the full-length proteins for the different kinomes and analyzed their aggregation properties. Surprisingly, the calculated average aggregation propensity Na4vSS was positive in the AGGRESCAN scale in all cases, which suggests a certain intrinsic propensity to aggregate for these domain sequences (**Figure 2A**). Na4vSS of 1.56, 1.42, 1.03, and 0.75 were calculated for yeast, fly, mouse, and human kinomes, respectively. Na4vSS values reflect the average propensity of all the proteins in a dataset. To compare the distribution of domains displaying positive aggregation propensity in the different species, relative to proteins in the Swiss-Prot database, we binned Na4vSS values into 100 groups and calculated the deviation between the human and the rest of kinomes for bins in which Na4vSS > 0.

$$d_{(x-\text{human})} = \sum_{i=1}^{N} F\left(Na4vSS_x\right) - F\left(Na4vSS_{\text{human}}\right) \qquad (1)$$

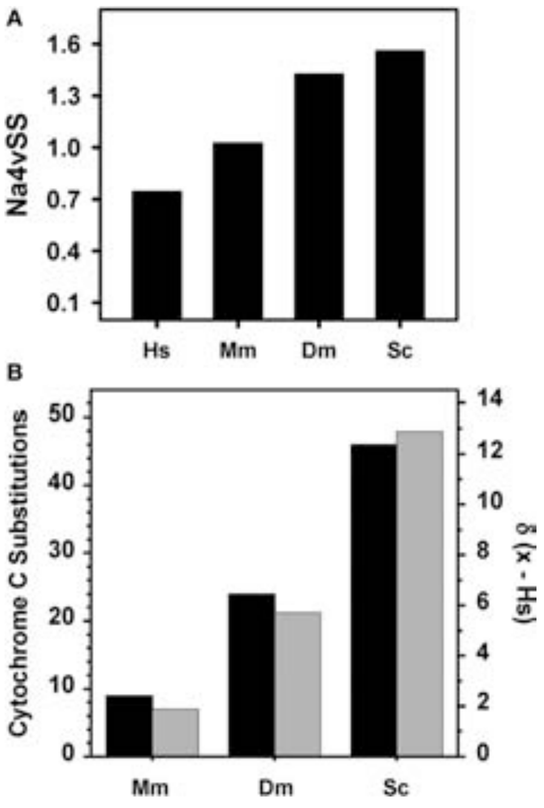


FIGURE 2 | Relationship between organism complexity and aggregation properties of kinase domains. (A) Average aggregation propensity (Na4vSS) for the complete dataset of kinome domains of human (*Hs*), mouse (*Mm*), fruit fly (*Dm*), and budding yeast (*Sc*). **(B)** Deviation in the distribution of aggregation prone domains (Na4vSS > 0) between human and the rest of species (gray, calculated according to Equation 1) compared with the number of amino acid differences in cytochrome c between human and other species (black).

where *F(Na4vSS$_x$)* corresponds to the frequency of this bin in the organism *x* and *F(Na4vSS$_{\mathrm{human}}$)* is its frequency in the human kinome.

The calculated deviations match well with the evolutive distances in the phylogenetic tree of cytochrome c (Dayhoff et al., 1972) (**Figure 2B**). Therefore, for kinase domains, it appears that aggregation propensity decreases as we ascend in the evolutionary scale.

We explored the reasons for the different aggregation propensities observed in the kinase domains of different species. The frequency of aggregating peaks NnHS is approximately four in all species (**Figure 3A**). This value is lower in yeast than in humans and therefore it cannot account for the observed differences in overall aggregation propensity. In contrast, the THSAr values follow the trend observed for Na4vSS, indicating that despite sharing similar number of aggregating peaks, the aggregation potency of these regions decreases with organism complexity (**Figure 3B**). This became more obvious when we compared the cumulative THSAr frequencies in distant organisms, yeast and human (**Figure 3C**). The 25% of the human kinase domains have a low THSAr (<0.1) in contrast to 5% of yeast domains. On the contrary, 20% of yeast domains display a high THSAr value (>0.15) while only 10% of human domains are included in this set. A similar, trend is observed for AATr values, yet another measure of the aggregation propensity of the sequence (**Figures 3D,E**).

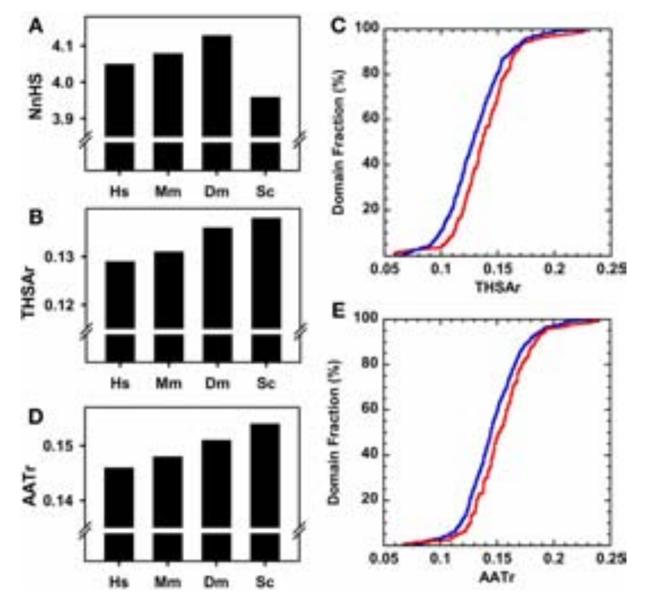


FIGURE 3 | Aggregation properties of complete kinase domain datasets of different organisms. (A) Normalized number of Hot-spots (NnHS). **(B)** Total Hot-spot area per residue (THSAr). **(C)** Distribution of the THSAr value along the whole dataset. **(D)** Area of the aggregation profile above the Hot-Spot Threshold per Residue (AATr). **(E)** Distribution of the AATr value over the complete kinase domain dataset. In **(A)**, **(B)**, and **(D)**: human (*Hs*), mouse (*Mm*), fruit fly (*Dm*), and budding yeast (*Sc*). In **(C)** and **(E)**: the blue line corresponds to the human dataset and the red line to the yeast kinome.

To see whether the differences in aggregation propensity of human and yeast sequences might result from an amino acid compositional bias in these species, we compared the amino acid content of yeast and human kinase domains with the average composition of the proteins deposited in Swiss-Prot (**Figure 4A**). Following the trend described above, both human and yeast kinase domains are, on the average, enriched in residues with high β-aggregation propensity (C, F, I, L, N, Q, V, and Y) and depleted in residues with low β-aggregation propensity (A, G, H, K, P, and R) (Tartaglia et al., 2005). However, these trends are more evident in yeast domains (**Figure 4B**), in agreement with their overall higher predicted aggregation properties.

AGGREGATION PROPERTIES OF THE HUMAN KINOME

We addressed how the aggregation properties of complete human kinase proteins and their domains compare to those of folded proteins. We used the SCOP-derived database ASTRAL40 (Chandonia et al., 2004) and randomly selected 500 sequences in order to obtain a database similar in size to the human kinome (508 domains and 497 proteins). Proteins in the ASTRAL40 database display higher Na4vSS values than the full-length human kinase protein set (**Figure 5A**). In addition, on the average, the frequency of aggregation-promoting regions is lower in human kinases than in the ASTRAL40 dataset (**Figure 5B**). Morevover, kinases tend to have less effective aggregating regions than the selected set of human folded proteins (**Figures 5C,D**). If we only consider the kinase domains in these proteins we observe the opposite trend. The 60% of kinase domains display positive Na4vSS, in contrast to the 22% of proteins in the ASTRAL40 dataset. In addition, only 2% of kinase domains have high-predicted solubility (Na4vSS < −10), whereas the 30% of folded human proteins display this property.

When we compare human kinase domains with the correspondent full-length proteins, their aggregation properties appear to be strikingly different. For all the parameters, their cumulative frequencies run parallel but always with higher values for the domains alone (**Figures 5A–D**). This arises from a higher density of aggregating peaks in the domains, displaying also higher potency, compared to the complete protein in which they reside,

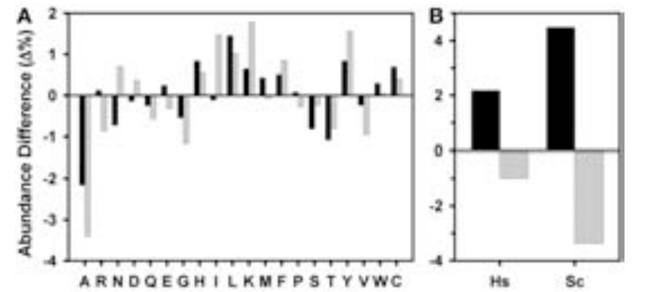


FIGURE 4 | Amino acid composition of kinase domains. (A) Differential amino acid composition of the human (black) and yeast (gray) kinase domains and **(B)** differential proportion of residues with high (black) and low (gray) β-aggregation propensity in human (Hs) and yeast (Sc) kinase domains, both relative to the amino acid composition of the Swiss-Prot database ensemble (release 2012_07).

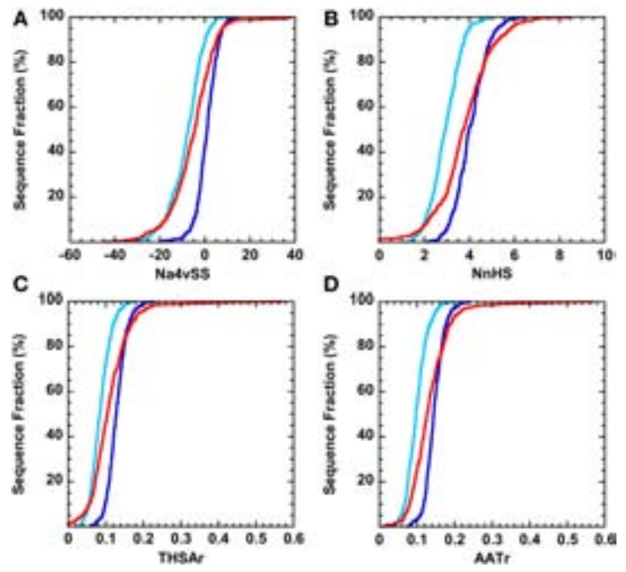


FIGURE 5 | Distribution of aggregation properties over the complete datasets of human kinase proteins (light blue) and domains (dark blue). The red line corresponds to the distribution for a randomized selection of 500 sequences belonging to the Astral40 dataset. **(A)** Average aggregation propensity (Na4vSS). **(B)** Normalized number of Hot-spots (NnHS). **(C)** Total Hot-spot area per residue (THSAr). **(D)** Area of the aggregation profile above the Hot-Spot Threshold per Residue (AATr).

as shown in **Figures 6A–C**. It has been shown that intrinsically unstructured proteins (IUPs) and segments are inherently less aggregation-prone than globular proteins (de Groot and Ventura, 2010). Therefore, we wondered if there is any relationship between Na4vSS values and the presence of disordered regions in full-length human proteins. With this aim we calculated the intrinsic unfolding propensity of the 10% protein kinases with the lowest and highest associated Na4vSS values, respectively, using the FoldIndex algorithm. Proteins displaying low intrinsic aggregation propensity are enriched in disordered regions (**Figure 6D**). This behavior is common to all kinomes, as illustrated in **Figure 6D** where we compare human and yeast proteins.

AMYLOIDOGENIC REGIONS IN HUMAN KINASE DOMAINS

AGGRESCAN is a good predictor of intracellular aggregation propensity and of the regions driving this process; however it is not aimed to identify regions that self-assemble into ordered amyloids, albeit aggregation-prone and amyloidogenic regions coincide in many cases (Rousseau et al., 2006a). To extend our analysis, we have also evaluated the presence of patterns corresponding to hexa-peptides that form amyloid-like fibrils as deduced from experimental studies by de la Paz and Serrano (de la Paz and Serrano, 2004). The patterns were detected using ScanProsite (http://prosite.expasy.org/scanprosite/). One thousand and twenty-nine hits were obtained in 435 out of the 508 active domain sequences that include the human kinome dataset, thus indicating that the presence of sequences able to form amyloid structures, if exposed to solvent, is frequent in kinase domains, many of the sequences containing more than one amyloidogenic stretch.

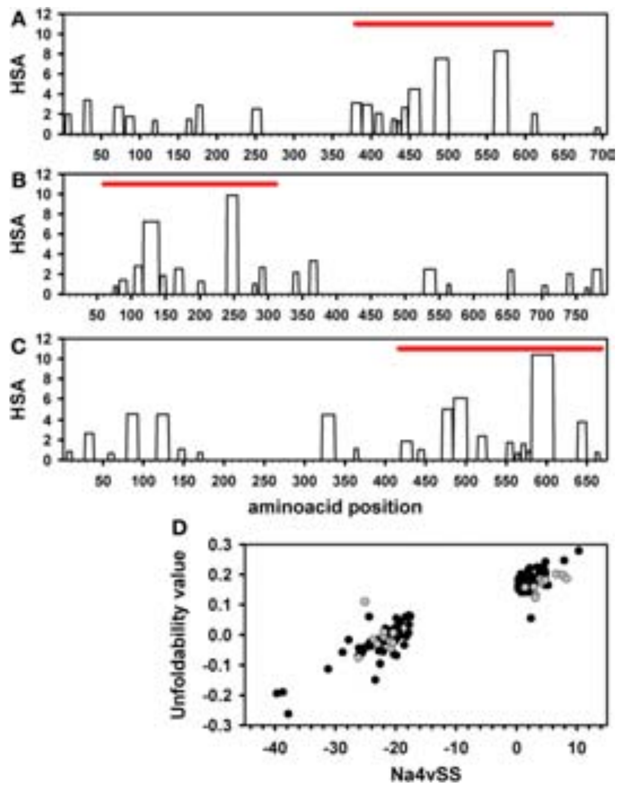


FIGURE 6 | Aggregation prone-regions and intrinsic disorder in kinases. AGGRESCAN Hot-Spot Area (HSA) profile for individual kinase proteins: **(A)** PKCt (AGC), **(B)** MARK1 (CAMK), and **(C)** BMX (TK). The position of the corresponding kinase domain is represented by a red bar. **(D)** Intrinsical unfolding propensity of the 10% kinase proteins with the lowest and highest Na4vSS values in the human (black) and yeast (gray) datasets. Intrinsical unfolding propensity predictions were performed using the FoldIndex algorithm.

AGGREGATION AND AMYLOIDOGENIC PROPERTIES OF HUMAN KINASE GROUPS

Typical protein kinase domains share a common catalytic core consisting of a small, mostly β-sheet, N-terminal subdomain and a larger, mostly -helical, C-terminal subdomain (Taylor and Radzio-Andzelm, 1994). Despite sharing a common structural core, typical human protein kinases show significant sequential divergence and can be classified in 119 different families corresponding to 9 major groups (Manning et al., 2002b), excluding atypical protein kinases, which do no have structural similarity to the rest of protein kinases. The proteins in a given group are sequentially related. Therefore, we explored if the different groups have characteristic aggregative and amyloidogenic properties. For statistical purposes we analyzed only those groups containing at least 40 sequences: AGC (67 domains), CAMK (69 domains), CMGC (64 domains), Other (80 domains), STE (47 domains), Tyrosine kinase (93 domains), and Tyrosine kinase-like (41 domains). This subset accounts for 89% of human protein kinase sequences. In **Figure 7** we compare the Na4vSS values for the kinase domains and the complete proteins. With the exception of those in the heterogeneous group Other, the

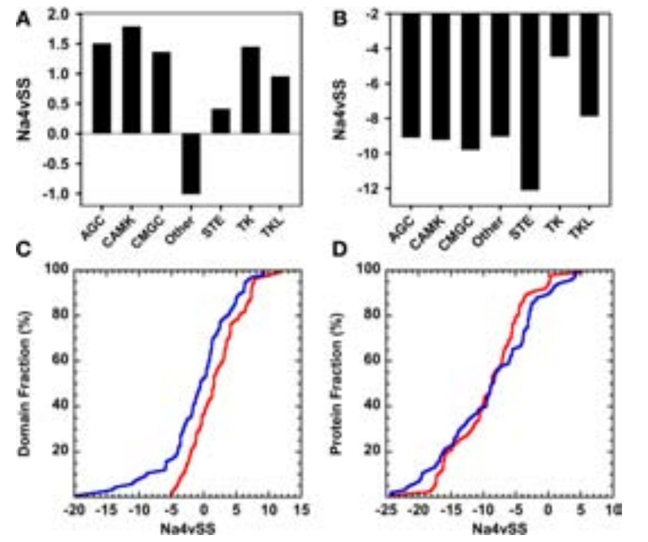


FIGURE 7 | Aggregation properties of the human kinase proteins and domains clustered according to their corresponding group (only groups composed of >40 sequences are considered). (A) and **(B)** Average aggregation propensity (Na4vSS) of the sets corresponding to each group considering only domains or whole proteins, respectively. **(C)** and **(D)** Distribution of the average aggregation propensity (Na4vSS) across the sets corresponding to the domain groups that show the most opposed behavior: Other (blue) and CAMK (red). Distributions are shown for kinase domains only **(C)** and for whole proteins **(D)**.

domains in the rest of the groups display positive average aggregation propensity (**Figure 7A**), with CAMK and STE groups displaying the highest and lowest aggregation propensities, respectively. These differences might reflect functional constrains, since kinases in the CAMK group tend to form complexes by auto-association whereas STE kinases act transducing signals from the surface of the cell to the nucleus and should be inherently soluble. All full-length proteins display negative Na4vSS values (**Figure 7B**). The analysis of Na4vSS cumulative frequencies in different groups does not show any correlation between the aggregation propensity of the domain and that of the protein within it is included (**Figures 7C,D**).

We used the structural alignment reported by Scheeff and Bourne (Scheeff and Bourne, 2005) to map the above mentioned amyloidogenic hexa-peptides over the three-dimensional structures of kinase domains. Thirteen non-redudant structures, representative of the domains in the different human groups, were selected for the analysis (**Figure 8**). About 90% of the predicted amyloidogenic regions overlap with secondary structure elements in the kinase domains. These stretches are sequence dependent and, therefore, they map in different regions in the different domains; however the region comprising the well conserved β-hairpin formed by β-sheets 4 and 5 at the N-terminal subdomain seems to be specially amyloidogenic. Importantly, the hydrophobic side chains in these two β-sheets are usually buried inside the structure, thus preventing the aggregation of the native domain under physiological conditions. The amyloidogenic regions in the mitogen-activated protein kinase-activated protein kinase 2, belonging to the CAMK group, which include β-sheets 4 and 5, are shown in **Figure 9** over the domain structure.

AGGREGATION PROPERTIES OF PROTEIN KINASE GROUPS IN DIFFERENT SPECIES

To test if the observed relationship between domains aggregation propensity and organisms complexity is maintained at the group level, we compared the Na4vSS of human kinase groups with those in yeast, fly, and mouse (**Figure 10**). It is important to note that the human kinome contains two and five times more kinases than the kinomes of fly and yeast, respectively. Therefore, the number of sequences in the considered groups differs between organisms. A strict correlation between aggregation propensity and organism complexity is seen for domains belonging to kinases in the AGC, CMG, Other, and TK Tyrosine kinase groups. However, deviation from this principle is observed in the CAMK group in which yeast domains appear to be less aggregation-prone than those in multicellular organisms. This might reflect the fact that three out of the seven yeast-specific kinase subfamilies and 30% of the specific yeast kinase genes belong to the CAMK group (Manning et al., 2002a). They mediate essentially unicellular-specific functions (Ball et al., 2000), including osmotic and other stress responses, which require a significant degree of solubility.

We addressed whether the location of catalytically important residues might constraint the evolution of aggregation propensity. To this aim, we analyzed the presence of conserved catalytic residues inside aggregation-prone stretches in representative proteins corresponding to different groups in human, fly, and yeast (**Figure 11**). The results indicate that, independently of the protein group and species, a significant amount of catalytic residues are embedded in aggregation-prone regions, likely limiting the evolution of these regions toward more soluble sequences.

DISCUSSION

Computational predictions of the aggregation propensities of proteomes in different organisms suggest that minimization of protein aggregation may act as an important evolutionary constrain, shaping proteins and cellular machineries (Castillo et al., 2011). This negative selection pressure has been suggested to modulate the aggregation propensity of protein sequences according to their biological context (Rousseau et al., 2006b; Tartaglia et al., 2007; Monsellier et al., 2008; Castillo and Ventura, 2009; Tartaglia and Vendruscolo, 2009; de Groot and Ventura, 2010). In this line, a previous analysis of the overall aggregation tendency of complete proteomes, including those of yeast, fly, mouse, and human, has shown that this value decreases with increasing organism complexity and longevity (Tartaglia et al., 2005). This trend results from the fact that higher organisms contain fewer sequences with high aggregation propensity, but specially because their proteomes contain a higher proportion of IUPs. In addition, the force of natural selection against the aggregation of a given protein depends on the selective contribution of this protein to the organism fitness (Chen and Dokholyan, 2008; Villar-Pique et al., 2012). Therefore, despite the analysis of complete proteomes has contributed significantly to our present understanding on how aggregation modulates protein

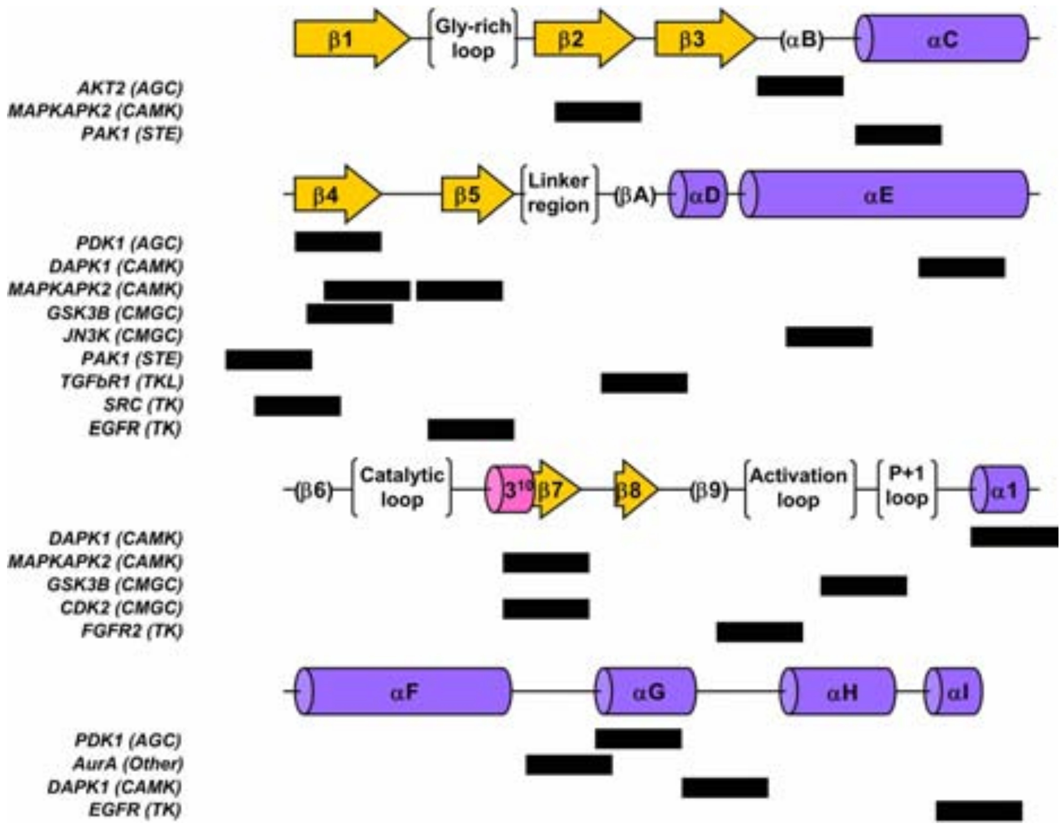


FIGURE 8 | Location of the detected amyloidogenic patterns within the global kinase fold structural elements for a representative set of the different groups of human kinase domains. Structural elements are depicted following the nomenclature in (Scheeff and Bourne, 2005); β-sheets are shown as arrows, -helices as cylinders and other structural or functional features are shown in parenthesis or brackets. Black boxes denote the location of the hexapeptidic pattern found for the corresponding inline domain.

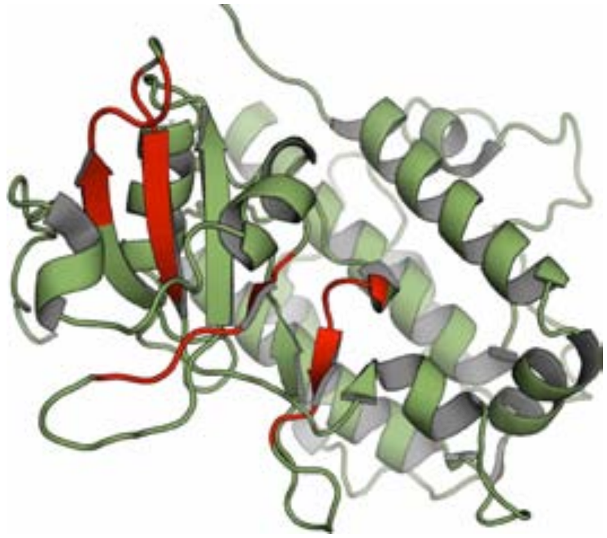


FIGURE 9 | Location of the detected amyloidogenic patterns on the structure of MAPKAPK2 belonging to the CAMK group (PDB 1kwp), amino acid stretches corresponding to the detected patterns are shown in red.

sequences and structures, the datasets used in these studies comprise proteins displaying divergent structures, from globular to unfolded, and functions, from essential to spurious, and thus under different evolutionary pressures.

Most evolutionary studies rely on the analysis of sequential or structural differences in a given protein family or super-family. By analogy, the study of the aggregation properties of a homogeneous, but relatively large, group of proteins sharing the same fold and function in different organisms might help to confirm the constraint aggregation exerts in protein evolution. The cellular role of the proteins in this set should be important enough to be the subject of a significant selection against their aggregation, in order to avoid a deleterious loss of function. The superfamily of protein kinases is perfectly suited for this purpose. We could confirm here that, on the average, the kinase domains of more complex and longer living organisms tend to have lower aggregation propensities and less potent aggregating peaks in their sequence.

Our predictions argue that, independently of the considered organism and group, kinase domains have significant aggregation tendency. Accordingly, human domains display higher aggregation propensity than the selected ensemble of folded proteins. Interestingly enough, it has been recently shown that 60% of

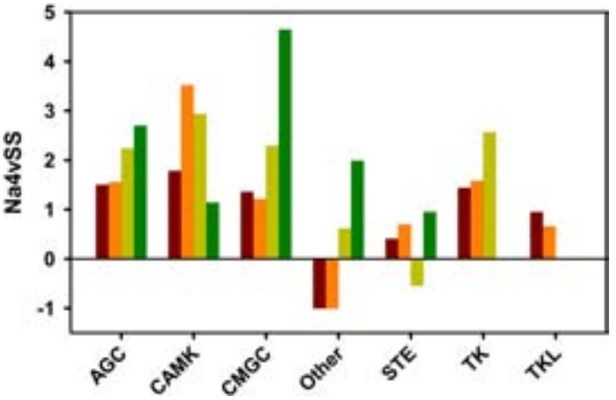


FIGURE 10 | Average aggregation propensity (Na4vSS) for discrete domain datasets of kinase groups in different organisms: human (brown), mouse (orange), fruit fly (light green), and budding yeast (dark green). Only the groups considered previously in the human kinome analysis are represented.

human kinases are clients of the HSP90 chaperone and that their binding determinants are located in the kinase domain (Taipale et al., 2012). In these cases, inhibition of the chaperone binding activity results in dissociation of the kinase domain and leads to aggregation, supporting the view that kinase domains are intrinsically aggregation-prone. This property might explain why, when recombinantly expressed in bacteria, kinase domains accumulate as misfolded and insoluble aggregates in many cases (Benetti et al., 1998; Marin et al., 2010) since prokaryotic HSP90 does not chaperone kinases (Buchner, 2010).

Despite the recognized cytoxicity of amyloid aggregates, it is now clear that amyloidogenic sequences are ubiquitous in all the proteomes (Castillo et al., 2011), supporting the view that most polypeptides share the potential to form amyloid-assemblies (Dobson, 2004). Accordingly, we found that 85% of the human kinase domains include at least one amyloidogenic stretch and in many cases several of them. These data are in agreement with the observation that, *in vitro*, the population of partially unfolded states results in amyloid aggregation in different, unrelated, kinase domains (Damaschun et al., 2000; Agocs et al., 2010, 2012; Georgescauld et al., 2011). The mapping of these amyloidogenic regions on the three-dimensional structures of catalytic domains belonging to different human kinase groups indicates that, in most cases, they overlap with regular elements of secondary structure, providing support to the hypothesis that the molecular determinants responsible for amyloid formation coincide with those maintaining the native structure of proteins (Sabate et al., 2012). In other words, kinase domains cannot avoid the presence of amyloid regions simply because they need them to fold into compact and stable globular conformations. Moreover, aggregation-prone regions overlap with conserved functional residues, suggesting that the selection for active conformations during evolution constrains the pressure to attain more soluble sequences.

In the native state, the side-chains of amyloidogenic sequences are essentially protected from the solvent inside the three-dimensional structure of kinase domains. However, mutations or environmental conditions that destabilize the functional conformation would result in their partial or total exposition to solvent, allowing them to nucleate amyloid self-assembly, a mechanism common to several proteins involved in human conformational disorders (Monsellier and Chiti, 2007). This model would explain why certain kinases, such as c-Src and EGFR, associate transiently with HSP90 chaperone during maturation, where these regions are likely exposed to solvent, but do not bind to this chaperone when they are fully folded (Xu et al., 1999, 2005).

Despite the effect of selective pressure could be clearly traced for kinase domains, full-length kinases display, on the average, low aggregation propensity, independently of the considered species. Our analysis indicates that the presence of intrinsically unstructured regions in non-catalytic domains is, at least in part, responsible for the low sequential aggregation propensity of kinases. It has been recently shown that intrinsically disordered protein sequences traslationally fused to globular proteins act as entropic bristles, providing them solubility by creating both a large favorable surface area for water interactions and large excluded volumes around the partner (Santner et al., 2012). In kinases, this effect seems likely to act as a compensatory mechanism for the obligatory presence of aggregation-prone sequences in the catalytic domain.

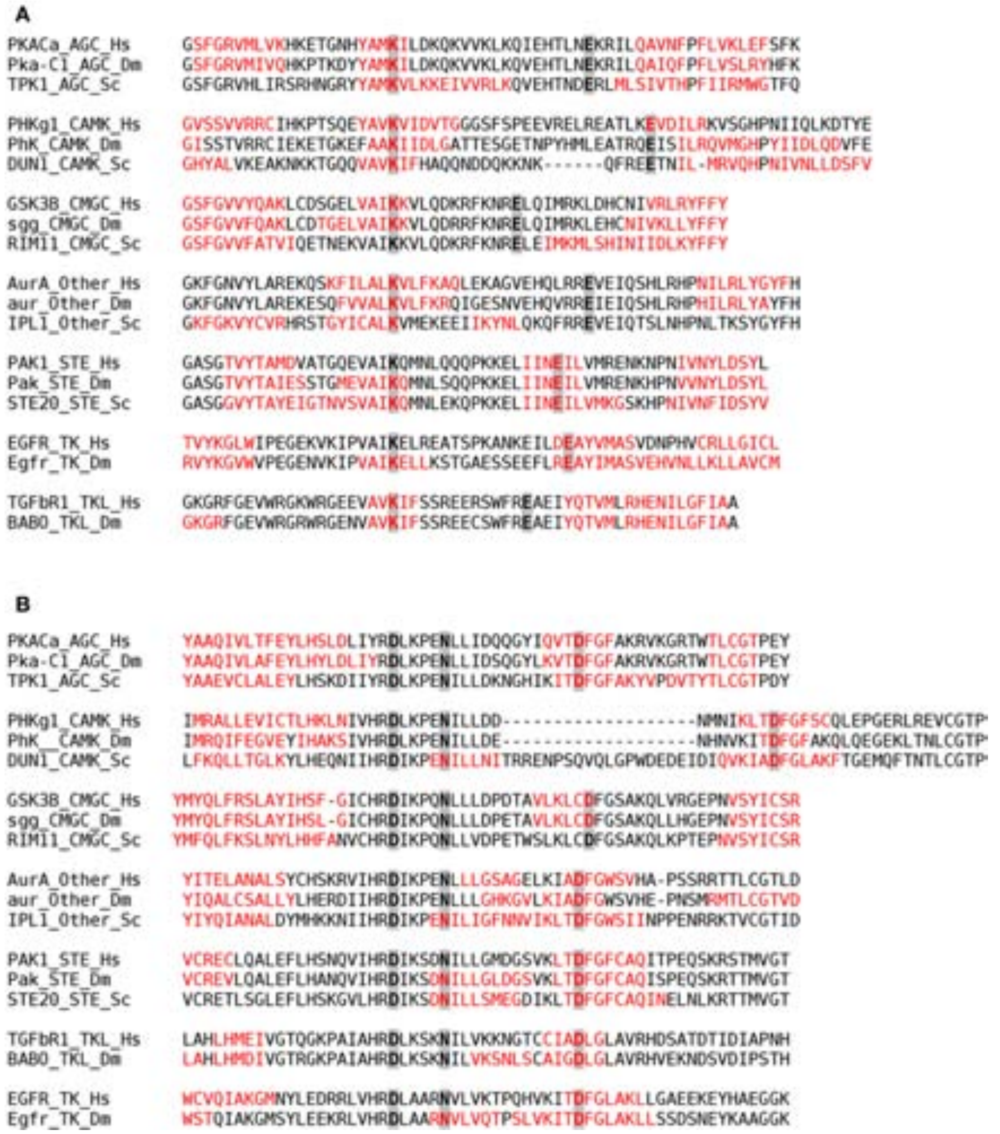
Apart from structural and catalytic constrains, there is an alternative and, although speculative, more interesting explanation for the calculated aggregation propensity of kinase domains: that it serves for functional purposes. It has been shown that despite kinases can rapidly evolve away from chaperone assisted folding they have generally not done so. This suggests that kinases might specifically exploit association with the chaperone machinery as a means of regulation, in such a way that the inherent instability of some kinases might in fact be employed as a mechanism to provide fidelity in regulatory cascades (Taipale et al., 2012). The exposition of pre-existing aggregation-prone regions able to bind the chaperone upon partial unfolding would likely play an important role in this mechanism.

Despite the present work illustrates the utility of prediction algorithms to provide insights on the relationship between protein aggregation and sequence evolution, it should be noted that at the present moment these algorithms do not allow to evaluate the impact of post-translational modifications on aggregation propensity. It is clear that incorporating these modifications, which might promote extensive structural rearrangements, in the calculations would result in a more realistic readout on the evolutive constrains imposed by protein aggregation and that we should work toward this objective.

MATERIALS AND METHODS

DATASET CURATION

Kinase Domains and Entire Protein sequences where retrieved from the Kinbase Database (http://kinase.com/kinbase/) for *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, and *Saccharomyces cerevisiae*. Domains or proteins with "not available" ("N/A") sequence or containing the non-amino acidic



```
A

PKACa_AGC_Hs      GSFGRVMLVKHKETGNHYAMKILDKQKVVKLKQIEHTLNEKRILQAVNFPFLVKLEFSFK
Pka-C1_AGC_Dm     GSFGRVMIVQHKPTKDYYAMKILDKQKVVKLKQVEHTLNEKRILQAIQFPFLVSLRYHFK
TPK1_AGC_Sc       GSFGRVHLIRSRHNGRYYAMKVLKKEIVVRLKQVEHTNDERLMLSIVTHPFIIRMWGTFQ

PHKg1_CAMK_Hs     GVSSVVRRCIHKPTSQEYAVKVIDVTGGGSFSPEEVRELREATLKEVDILRKVSGHPNIIQLKDTYE
PhK_CAMK_Dm       GISSTVRRCIEKETGKEFAAKIIDLGATTESGETNPYHMLEATRQEISILRQVMGHPYIIDLQDVFE
DUN1_CAMK_Sc      GHYALVKEAKNKKTGQQVAVKIFHAQQNDDQKKNK------QFREETNIL-MRVQHPNIVNLLDSFV

GSK3B_CMGC_Hs     GSFGVVYQAKLCDSGELVAIKKVLQDKRFKNRELQIMRKLDHCNIVRLRYFFY
sgg_CMGC_Dm       GSFGVVFQAKLCDTGELVAIKKVLQDRRFKNRELQIMRKLEHCNIVKLLYFFY
RIM11_CMGC_Sc     GSFGVVFATVIQETNEKVAIKKVLQDKRFKNRELEIMKMLSHINIIDLKYFFY

AurA_Other_Hs     GKFGNVYLAREKQSKFILALKVLFKAQLEKAGVEHQLRREVEIQSHLRHPNILRLYGYFH
aur_Other_Dm      GKFGNVYLAREKESQFVVALKVLFKRQIGESNVEHQVRREIEIQSHLRHPHILRLYAYFH
IPL1_Other_Sc     GKFGKVYCVRHRSTGYICALKVMEKEEIIKYNLQKQFRREVEIQTSLNHPNLTKSYGYFH

PAK1_STE_Hs       GASGTVYTAMDVATGQEVAIKQMNLQQQPKKELIINEILVMRENKNPNIVNYLDSYL
Pak_STE_Dm        GASGTVYTAIESSTGMEVAIKQMNLSQQPKKELIINEILVMRENKHPNVVNYLDSYL
STE20_STE_Sc      GASGGVYTAYEIGTNVSVAIKQMNLEKQPKKELIINEILVMKGSKHPNIVNFIDSYV

EGFR_TK_Hs        TVYKGLWIPEGEKVKIPVAIKELREATSPKANKEILDEAYVMASVDNPHVCRLLGICL
Egfr_TK_Dm        RVYKGVWVPEGENVKIPVAIKELLKSTGAESSEEFLREAYIMASVEHVNLLKLLAVCM

TGFbR1_TKL_Hs     GKGRFGEVWRGKWRGEEVAVKIFSSREERSWFREAEIYQTVMLRHENILGFIAA
BABO_TKL_Dm       GKGRFGEVWRGRWRGENVAVKIFSSREECSWFREAEIYQTVMLRHENILGFIAA

B

PKACa_AGC_Hs      YAAQIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWTLCGTPEY
Pka-C1_AGC_Dm     YAAQIVLAFEYLHYLDLIYRDLKPENLLIDSQGYLKVTDFGFAKRVKGRTWTLCGTPEY
TPK1_AGC_Sc       YAAEVCLALEYLHSKDIIYRDLKPENILLDKNGHIKITDFGFAKYVPDVTYTLCGTPDY

PHKg1_CAMK_Hs     IMRALLEVICTLHKLNIVHRDLKPENILLDD-------------------NMNIKLTDFGFSCQLEPGERLREVCGTP
PhK__CAMK_Dm      IMRQIFEGVEYIHAKSIVHRDLKPENILLDE-------------------NHNVKITDFGFAKQLQEGEKLTNLCGTP
DUN1_CAMK_Sc      LFKQLLTGLKYLHEQNIIHRDIKPENILLNITRRENPSQVQLGPWDEDEIDIQVKIADFGLAKFTGEMQFTNTLCGTP

GSK3B_CMGC_Hs     YMYQLFRSLAYIHSF-GICHRDIKPQNLLLDPDTAVLKLCDFGSAKQLVRGEPNVSYICSR
sgg_CMGC_Dm       YMYQLFRSLAYIHSL-GICHRDIKPQNLLLDPETAVLKLCDFGSAKQLLHGEPNVSYICSR
RIM11_CMGC_Sc     YMFQLFKSLNYLHHFANVCHRDIKPQNLLVDPETWSLKLCDFGSAKQLKPTEPNVSYICSR

AurA_Other_Hs     YITELANALSYCHSKRVIHRDIKPENLLLGSAGELKIADFGWSVHA-PSSRRTTLCGTLD
aur_Other_Dm      YIQALCSALLYLHERDIIHRDIKPENLLLGHKGVLKIADFGWSVHE-PNSMRMTLCGTVD
IPL1_Other_Sc     YIYQIANALDYMHKKNIIHRDIKPENILIGFNNVIKLTDFGWSIINPPENRRKTVCGTID

PAK1_STE_Hs       VCRECLQALEFLHSNQVIHRDIKSDNILLGMDGSVKLTDFGFCAQITPEQSKRSTMVGT
Pak_STE_Dm        VCREVLQALEFLHANQVIHRDIKSDNILLGLDGSVKLTDFGFCAQISPEQSKRTTMVGT
STE20_STE_Sc      VCRETLSGLEFLHSKGVLHRDIKSDNILLSMEGDIKLTDFGFCAQINELNLKRTTMVGT

TGFbR1_TKL_Hs     LAHLHMEIVGTQGKPAIAHRDLKSKNILVKKNGTCCIADLGLAVRHDSATDTIDIAPNH
BABO_TKL_Dm       LAHLHMDIVGTRGKPAIAHRDLKSKNILVKSNLSCAIGDLGLAVRHVEKNDSVDIPSTH

EGFR_TK_Hs        WCVQIAKGMNYLEDRRLVHRDLAARNVLVKTPQHVKITDFGLAKLLGAEEKEYHAEGGK
Egfr_TK_Dm        WSTQIAKGMSYLEEKRLVHRDLAARNVLVQTPSLVKITDFGLAKLLSSDSNEYKAAGGK
```

FIGURE 11 | Aggregation-prone regions detected by AGGRESCAN within the 20 amino acids flanking the Nter and Cter sides of the first (A) and second (B) regions containing conserved catalytic residues in kinase domains. Aggregation-prone stretches are shown in red in the alignment of representative proteins corresponding to different classes (AGC, CAMK, CMGC, Other, STE, TK, and TKL) in human (Hs), fruit fly (Dm), and yeast (Sc). The conserved catalytic residues are highlighted in gray.

character "∗" were deleted. Whenever the non-amino acidic character "X" was found, it was substituted by alanine ("A"), unless more than three consecutive "X" were found, in which case the sequence was deleted. Single spaces found within sequences were suppressed.

Sequences in Kinase Domain datasets were verified to correspond to sequential regions of the Entire Protein datasets. Those domains without a correspondent sequence in the Entire Protein dataset were deleted. In the same way, full-length proteins without an associated kinase domain were also suppressed.

The Kinase Domain and Entire Protein datasets obtained in this way were further subdivided according to the different groups of kinases present in each species.

Due to AGGRESCAN technical limitations, parameters for sequences larger than 2000 amino acids could not be computed and those sequences and their corresponding domains were deleted from the datasets.

AGGRESCAN CALCULATIONS

AGGRESCAN (http://bioinf.uab.es/aggrescan/) was employed, with default settings, in order to compute parameters indicative of aggregation properties for every single sequence, as well as overall values for complete datasets. In order to be able to compare the large datasets of this study, only normalized values were considered (**Figure 1**). For further information about the calculation of AGGRESCAN parameters and its applications the reader is referred to AGGRESCAN Help File (accessible from the server front page) and to the published tutorials (de Groot et al., 2012).

COMPOSITION OF KINASE DOMAINS

The frequencies of occurrence of each amino acid were calculated for the complete kinase domains datasets of human and budding yeast. In order to look for compositional biases in these sets, frequency differences were calculated relative to the occurrence of each amino acid in the complete Swiss-Prot database release 2012_07.

ANALYSIS OF THE HUMAN KINOME

In order to be able to set up general comparisons between the human kinase proteins and its isolated domains, a reference dataset was defined by randomly retrieving, using a randomizing function, 500 sequences from the subset of sequences with less than 40% identity of the ASTRAL Compendium (Astral40), whose AGGRESCAN parameters were also calculated.

DETECTION OF AMYLOIDOGENIC PATTERNS

The human kinase domains dataset was explored to identify amyloidogenic patterns by using the amyloidogenic signature {P}-{PKRHW}-[VLSWFNQ]-[ILTYWFN]-[FIY]-{PKRH} defined in (de la Paz and Serrano, 2004) using the ScanProsite web server.

INTRINSICAL UNFOLDING PREDICTION

The intrinsical unfolding potential of selected subsets of the human and yeast kinase domains datasets was computed using the FoldIndex webserver with default settings.

ACKNOWLEDGMENTS

This work was supported by BFU2010-14901 from Ministerio de Ciencia e Innovacion (MCISpain) and 2009-SGR from AGAUR (Generalitat de Catalunya). Ricardo Graña-Montes is beneficiary of predoctoral FPU AP 2009-0948 fellowship from the Ministerio de Educación-Spain. Salvador Ventura has been granted an ICREA ACADEMIA award.

REFERENCES

Agocs, G., Solymosi, K., Varga, A., Modos, K., Kellermayer, M., Zavodszky, P., et al. (2010). Recovery of functional enzyme from amyloid fibrils. *FEBS Lett.* 584, 1139–1142.

Agocs, G., Szabo, B. T., Kohler, G., and Osvath, S. (2012). Comparing the folding and misfolding energy landscapes of phosphoglycerate kinase. *Biophys. J.* 102, 2828–2834.

Ball, C. A., Dolinski, K., Dwight, S. S., Harris, M. A., Issel-Tarver, L., Kasarskis, A., et al. (2000). Integrating functional genomic information into the Saccharomyces genome database. *Nucleic Acids Res.* 28, 77–80.

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

Benetti, P. H., Kim, S. I., Chaillot, D., Canonge, M., Chardot, T., and Meunier, J. C. (1998). Expression and characterization of the recombinant catalytic subunit of casein kinase II from the yeast *Yarrowia lipolytica* in *Escherichia coli*. *Protein Expr. Purif.* 13, 283–290.

Buchner, J. (2010). Bacterial Hsp90–desperately seeking clients. *Mol. Microbiol.* 76, 540–544.

Castillo, V., Espargaro, A., Gordo, V., Vendrell, J., and Ventura, S. (2010). Deciphering the role of the thermodynamic and kinetic stabilities of SH3 domains on their aggregation inside bacteria. *Proteomics* 10, 4172–4185.

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

Castillo, V., and Ventura, S. (2009). Amyloidogenic regions and interaction surfaces overlap in globular proteins related to conformational diseases. *PLoS Comput. Biol.* 5:e1000476. doi: 10.1371/journal.pcbi.1000476

Chandonia, J. M., Hon, G., Walker, N. S., Lo Conte, L., Koehl, P., Levitt, M., et al. (2004). The ASTRAL Compendium in 2004. *Nucleic Acids Res.* 32, D189–D192.

Chen, Y., and Dokholyan, N. V. (2008). Natural selection against protein aggregation on self-interacting and essential proteins in yeast, fly, and worm. *Mol. Biol. Evol.* 25, 1530–1533.

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

Conchillo-Sole, O., de Groot, N. S., Aviles, F. X., Vendrell, J., Daura, X., and Ventura, S. (2007). AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics* 8:65. doi: 10.1186/1471-2105-8-65

Daggett, V., and Fersht, A. R. (2009). Protein folding and binding: moving into unchartered territory. *Curr. Opin. Struct. Biol.* 19, 1–2.

Damaschun, G., Damaschun, H., Fabian, H., Gast, K., Krober, R., Wieske, M., et al. (2000). Conversion of yeast phosphoglycerate kinase into amyloid-like structure. *Proteins* 39, 204–211.

Dayhoff, M. O., Park, C. M., and McLaughlin, P. J. (1972). *Atlas of Protein Sequence and Structure.* Silver Spring, MD: National Biomedical Research Foundation.

de Groot, N., Pallares, I., Aviles, F., Vendrell, J., and Ventura, S. (2005). Prediction of "hot spots" of aggregation in disease-linked polypeptides. *BMC Struct. Biol.* 5:18. doi: 10.1186/1472-6807-5-18

de Groot, N. S., Castillo, V., Grana-Montes, R., and Ventura, S. (2012). AGGRESCAN: method, application, and perspectives for drug design. *Methods Mol. Biol.* 819, 199–220.

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

de Groot, N. S., and Ventura, S. (2010). Protein aggregation profile of the bacterial cytosol. *PLoS ONE* 5:e9383. doi: 10.1371/journal.pone.0009383

de la Paz, M. L., and Serrano, L. (2004). Sequence determinants of amyloid fibril formation. *Proc. Natl. Acad. Sci. U.S.A.* 101, 87–92.

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.

Georgescauld, F., Sabate, R., Espargaro, A., Ventura, S., Chaignepain, S., Lacombe, M. L., et al. (2011). Aggregation of the neuroblastoma-associated mutant (S120G) of the human nucleoside diphosphate kinase-A/NM23-H1 into amyloid fibrils. *Naunyn-Schmiedeberg's Arch. Pharmacol.* 384, 373–381.

Hamodrakas, S. J. (2011). Protein aggregation and amyloid fibril formation prediction software from primary sequence: towards controlling the formation of bacterial inclusion bodies. *FEBS J.* 278, 2428–2435.

Ivanova, M. I., Thompson, M. J., and Eisenberg, D. (2006). A systematic screen of beta(2)-microglobulin and insulin for amyloid-like segments. *Proc. Natl. Acad. Sci. U.S.A.* 103, 4079–4082.

Jahn, T. R., and Radford, S. E. (2005). The yin and yang of protein folding. *FEBS J.* 272, 5962–5970.

Linding, R., Schymkowitz, J., Rousseau, F., Diella, F., and 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.

Manning, G., Plowman, G. D., Hunter, T., and Sudarsanam, S. (2002a). Evolution of protein kinase signaling from yeast to man. *Trends Biochem. Sci.* 27, 514–520.

Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002b). The protein kinase complement of the human genome. *Science* 298, 1912–1934.

Marin, V., Groveman, B. R., Qiao, H., Xu, J., Ali, M. K., Fang, X. Q., et al. (2010). Characterization of neuronal Src kinase purified from a bacterial expression system. *Protein Expr. Purif.* 74, 289–297.

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

Monsellier, E., Ramazzotti, M., Taddei, N., and Chiti, F. (2008). Aggregation propensity of the human proteome. *PLoS Comput. Biol.* 4:e1000199. doi: 10.1371/journal.pcbi.1000199

Nelson, R., and Eisenberg, D. (2006). Recent atomic models of amyloid fibril structure. *Curr. Opin. Struct. Biol.* 16, 260–265.

Rousseau, F., Schymkowitz, J., and Serrano, L. (2006a). Protein aggregation and amyloidosis: confusion of the kinds? *Curr. Opin. Struct. Biol.* 16, 118–126.

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

Sabate, R., Espargaro, A., Grana-Montes, R., Reverter, D., and Ventura, S. (2012). Native structure protects SUMO proteins from aggregation into amyloid fibrils. *Biomacromolecules* 13, 1916–1926.

Santner, A. A., Croy, C. H., Vasanwala, F. H., Uversky, V. N., Van, Y. Y., and Dunker, A. K. (2012). Sweeping away protein aggregation with entropic bristles: intrinsically disordered protein fusions enhance soluble expression. *Biochemistry* 51, 7250–7262.

Scheeff, E. D., and Bourne, P. E. (2005). Structural evolution of the protein kinase-like superfamily. *PLoS Comput. Biol.* 1:e49. doi: 10.1371/journal.pcbi.0010049

Taipale, M., Krykbaeva, I., Koeva, M., Kayatekin, C., Westover, K. D., Karras, G. I., et al. (2012). Quantitative analysis of hsp90-client interactions reveals principles of substrate recognition. *Cell* 150, 987–1001.

Tartaglia, G. G., and Caflisch, A. (2007). Computational analysis of the *S. cerevisiae* proteome reveals the function and cellular localization of the least and most amyloidogenic proteins. *Proteins* 68, 273–278.

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

Tartaglia, G. G., Pellarin, R., Cavalli, A., and Caflisch, A. (2005). Organism complexity anti-correlates with proteomic beta-aggregation propensity. *Protein Sci.* 14, 2735–2740.

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

Taylor, S. S., and Radzio-Andzelm, E. (1994). Three protein kinase structures define a common motif. *Structure* 2, 345–355.

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. U.S.A.* 101, 7258–7263.

Villar-Pique, A., de Groot, N. S., Sabate, R., Acebron, S. P., Celaya, G., Fernandez-Busquets, et al. (2012). The effect of amyloidogenic peptides on bacterial aging correlates with their intrinsic aggregation propensity. *J. Mol. Biol.* 421, 270–281.

Xu, W., Yuan, X., Xiang, Z., Mimnaugh, E., Marcu, M., and Neckers, L. (2005). Surface charge and hydrophobicity determine ErbB2 binding to the Hsp90 chaperone complex. *Nat. Struct. Mol. Biol.* 12, 120–126.

Xu, Y., Singer, M. A., and Lindquist, S. (1999). Maturation of the tyrosine kinase c-src as a kinase and as a substrate depends on the molecular chaperone Hsp90. *Proc. Natl. Acad. Sci. U.S.A.* 96, 109–114.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 August 2012; accepted: 31 October 2012; published online: 20 November 2012.

Citation: Graña-Montes R, Sant'Anna de Oliveira R and Ventura S (2012) Protein aggregation profile of the human kinome. Front. Physio. ***3****:438. doi: 10.3389/fphys.2012.00438*

This article was submitted to Frontiers in Systems Biology, a specialty of Frontiers in Physiology.

Research Topic:
From structural to molecular systems biology: experimental and computational approaches to unravel mechanisms of kinase activity regulation in cancer and neurodegeneration

Topic Editor:
Matteo Barberis,
Humboldt University Berlin, Germany;
Max Planck Institute for Molecular Genetics, Berlin, Germany

5.3.- The Relationship between Foldability and Aggregation Propensity

Disulfide bonds have a great impact on the structural stability of globular proteins and, as discussed in section 4.2, they may exert a protective role against deleterious aggregation. Nonetheless, in the case of proteins that possess multiple disulfide bonds, this covalent linkage usually has a strong influence on their folding process, already introduced in section 1.2. During their oxidative folding, disulfide-rich domains (DRDs) populate metastable intermediates causing their folding reactions to become particularly slow *in vitro*, compared to the rest of globular proteins, lasting tens of minutes, as shortest, even in the presence of disulfide catalysts. A series of evidences suggest that, although the oxidative folding of DRDs is assisted *in vivo* by different proteins (Frand et al. 2000; Denoncin & Collet 2013), partially folded disulfide intermediates may be populated in the endoplasmic reticulum (Jansens et al. 2002; Safavi-Hemami et al. 2012) for a significant amount of time. As well, even completely reduced and substantially unfolded DRDs may accumulate in the cytoplasm prior to their import to the mitochondrial intermembrane space (Fischer et al. 2013). The population of such states would put this kind of proteins at a high risk of aggregation if they happened to present APRs along their sequences. In fact, the deposition into amyloid aggregates of several small disulfide-rich proteins is known to be associated to several human disorders (Chiti & Dobson 2006).

In order to understand how DRDs avoid aberrant aggregation and to gain insights on the determinants of the different oxidative folding pathways that may be followed by these proteins, we have used different computational tools to analyze their conformational and aggregative properties. For that purpose, we employed a curated dataset, containing proteins representatives of all the different folds populated by DRDs (Cheek et al. 2006).

In first place, theoretical calculations show that the impact of disulfide disruption generally results in a large destabilization of their structure, with average destabilization values between 5 and 8 kcal·mol^{-1}. Considering that the stability of the native state of globular proteins, relative to the unfolded state, lays in the range of 5-20 kcal·mol^{-1} (Dill 1990), this indicates that reduction of DRDs disulfide bonds leads to the population of predominantly unstructured conformations. Despite such an observation may seem obvious, and indeed a large collapse of the native structure has been experimentally demonstrated for a plethora of disulfide-rich proteins (Li et al. 1995; Chang 1997), the consequences of complete disulfide disruption had not been addressed from a broad perspective for the ensemble of DRDs.

The calculated structural instability of DRDs upon disulfide disruption suggests these proteins may resemble IDPs in their fully reduced state. This hypothesis was evaluated by employing several computational tools for the prediction of protein disorder, yielding an important fraction of the DRDs predicted to be fully unstructured or to possess large disordered segments. However, in the amino acid scales employed by these algorithms, Cys is ranked as an order-promoting residue, probably because it is usually found forming disulfide bridges in the interior of globular proteins but not on the basis of its intrinsic physical-chemical properties (Marino & Gladyshev 2010). Due to their implicit enrichment in Cys residues, predictions performed on wt DRDs sequences might be inherently biased. In this sense, substitutions of Cys by amino acids of a similar small size, like Ala, or chemical structure, like Ser, might allow to better model the conformational properties of reduced DRDs. In fact, when the predictions are performed on the DRDs sequences incorporating such substitutions, the fraction of predicted disorder raises and a majority of these domains are predicted as fully disordered or as containing large disordered segments.

Since DRDs are predicted to be substantially unfolded in their reduced state, and considering they can significantly populate completely or partially reduced states *in vivo*, the presence of APRs in their sequences would result dangerous, because aggregation-prone stretches in DRDs are likely to become exposed for a sufficient amount of time so as to establish intermolecular contacts, then leading to aggregation. The analysis of the aggregation properties of DRDs reveals that these proteins are notably less aggregation-prone than a reference set of globular proteins, and actually share remarkable similarities with IDPs in their aggregation propensities. Additionally, DRDs also present a extremely low number of Hsp70 chaperone binding sites, with more than 80% of the sequences in the analyzed set devoid of them, thus exhibiting a proportion of binding sites even below that found for the IDPs reference set. The coincidence in conformational and aggregational properties between DRDs in their reduced state and IDPs suggests those characteristics may arise from similar compositional biases in their sequences, relative to the ensemble of globular proteins. Cys is, obviously, the most overrepresented amino acid in DRDs sequences, in contrast it is one of the most depleted residues in IDPs (Radivojac et al. 2007), such a feature accounts for the different conformational properties between DRDs in their oxidized state and IDPs. Dispensing Cys residues, the analysis of DRDs composition shows that, compared to globular proteins, they are mostly depleted in aliphatic amino acids while enriched in β-breaker (Pro and Gly) and some basic polar/charged residues (Asn and Lys). Therefore, in DRDs a similar compositional bias, with depletion in aggregation-prone residues combined with enrichment in amino acids difavouring aggregation, accounts for their low aggregation propensity as observed for IDPs; however, aromatic residues are not particularly underrepresented in DRDs as they are in IDPs likely its role in binding and, as well, IDPs are more homogeneously enriched in polar and charged residues of both acidic and basic character.

Despite their intrinsically low aggregation propensity, amyloidogenic stretches are still found in a small number of DRDs. Similarly to what is usually found in globular proteins (Tzotzos & Doig 2010), these stretches frequently map to regular secondary structure elements. In the case of DRDs, although amyloidogenic fragments map to all kinds of secondary structure elements, a majority of them is found within α-helices. This “hidding” of amyloidogenic stretches within α-helical structure, thus conformationally incompatible with the formation of amyloid-like structure, has been regarded as a strategy to avoid aberrant aggregation. Aside from these considerations, the residual presence of amyloid-promoting fragments in certain DRDs suggests they may serve for these DRDs to increase their folding efficiency. As introduced in section 1.2, there exist two major mechanisms followed by disulfide-rich proteins in their oxidative folding. In one of them, proteins fold though a limited succession of intermediates with native-like disulfide connectivities, while the other is characterized by the population of a highly heterogeneous ensemble of different cross-linked intermediates; between these two extremes, intermediate mechanisms are also possible. Attempts to predict the mechanisms by which a DRD would fold have proven unfruitful. We have assessed whether the differential aggregation properties within DRDs constitute a reflection of their foldability, and whether such a property may help to explain the oxidative folding of a particular disulfide-rich protein, by employing computational tools to analyze the intrinsic disorder and aggregation propensities of different DRDs whose folding mechanism have been elucidated. Proteins that fold through a complex ensemble of intermediates, like hirudin (Chatrenet & Chang 1993; Chang et al. 1995) or the module LA5 form the low-density lipoprotein receptor (Arias-Moreno et al. 2008), present significant predicted disorder and very low aggregation propensity, while disulfide-rich proteins folding through discrete native-like intermediates, like LDTI (Arolas et al. 2008; Pantoja-Uceda et al. 2009), are predicted to contain a low amount of disordered regions but an increased tendency to aggregate. Interestingly, the same trend is observed when comparing related DRDs with different oxidative folding mechanisms but sharing the same disulfide connectivity, like different conotoxins (Fuller et al. 2005), and even when they possess the same native topology, like TAP (Chang 1996) and BPTI (Weissman & Kim 1991) or proinsulin and IGF-1 (Guo et al. 2008). Since the forces driving both folding and aggregation overlap to a certain extent, these observations indicate that, in DRDs, their folding determinants coincide with APRs. Therefore, the oxidative folding mechanism a disulfide-rich protein will follow is mostly influenced by the potency of the folding determinants coded in their sequence. In DRDs where those folding determinants are strong enough, these guide the folding of the polypeptide and the intermediates are predominantly native-like, although DRDs folding determinants do not suffice to maintain the stability of the native structure and the presence of disulfide bridges is required. When such determinants are weak or absent, DRDs folding proceeds in a more stochastic way with disulfide bonds forming and reshuffling almost at random, within the restrictions imposed by the conformational flexibility of the polypeptide chain, until an appropriate connectivity is reached that stabilizes secondary structure propensities within the polypeptide.

The overlapping between APRs and folding determinants in DRDs would allow predicting the oxidative folding mechanism of disulfide-rich proteins. Those proteins exhibiting higher aggregation propensity, for which sufficiently potent APRs can be detected, are expected to exhibit a greater folding efficiency and, therefore, to fold through a BPTI-like mechanism. For those predicted as highly soluble, without significant aggregation-prone stretches, their intrinsic foldability would be low, so they can be expected to fold according to a hirudin-like mechanism. In order to evaluate whether prediction of aggregation propensity could serve as a tool to forecast the oxidative folding mechanism of DRDs, we have tested it with a novel DRD whose mechanism had not been resolved in advance. The carboxypeptidase inhibitor from *N. versicolor* (NvCI) is a disulfide-rich protein predicted to possess a low intrinsic disorder and substantial aggregation propensity, with two APRs detected within its sequence, one of which is strongly amyloidogenic. Therefore NvCI is expected to fold efficiently through an oxidative pathway with few native-like intermediates. Indeed, NvCI behaves as a rapidly disulfide-rich folder, achieving its native conformation *in vitro* in less than 8 hours, and only populates two major intermediates. These results confirm the predictive value of aggregation propensity in order to estimate the foldability of DRDs.

The analysis presented here for DRDs illustrates the widely observed competition between the propensity of a polypeptide to fold and to aggregate into non-functional conformations. Since the driving forces of both folding and aggregation are of a similar nature, the sequential determinants of these two processes cannot be completely disentangled. Therefore, although a strong selective pressure shapes protein sequences in order to reduce their aggregation propensity, the extent of such selection is constrained in globular proteins by the requirement to achieve efficiently their correct three-dimensional structure. Such requirement for foldability in globular proteins acts as a strong limitation because excessive selection against aggregation-prone sequences would compromise the acquisition of the native structure. Further evolutionary selection to prevent aggregation would require additional stabilizing elements, such as disulfide bonds or external cofactors in order to preserve the native conformation.

ANTIOXIDANTS & REDOX SIGNALING
Volume 21, Number 3, 2014
© Mary Ann Liebert, Inc.
DOI: 10.1089/ars.2013.5543

ORIGINAL RESEARCH COMMUNICATION

Association Between Foldability and Aggregation Propensity in Small Disulfide-Rich Proteins

Hugo Fraga,* Ricardo Graña-Montes,* Ricard Illa, Giovanni Covaleda, and Salvador Ventura

Abstract

Aims: Disulfide-rich domains (DRDs) are small proteins whose native structure is stabilized by the presence of covalent disulfide bonds. These domains are versatile and can perform a wide range of functions. Many of these domains readily unfold on disulfide bond reduction, suggesting that in the absence of covalent bonding they might display significant disorder. ***Results:*** Here, we analyzed the degree of disorder in 97 domains representative of the different DRDs families and demonstrate that, in terms of sequence, many of them can be classified as intrinsically disordered proteins (IDPs) or contain predicted disordered regions. The analysis of the aggregation propensity of these domains indicates that, similar to IDPs, their sequences are more soluble and have less aggregating regions than those of other globular domains, suggesting that they might have evolved to avoid aggregation after protein synthesis and before they can attain its compact and covalently linked native structure. ***Innovation and Conclusion:*** DRDs, which resemble IDPs in the reduced state and become globular when their disulfide bonds are formed, illustrate the link between protein folding and aggregation propensities and how these two properties cannot be easily dissociated, determining the main traits of the folding routes followed by these small proteins to attain their native oxidized states. *Antioxid. Redox Signal.* 21, 368–383.

Introduction

LACKING HYDROPHOBIC CORES and stable secondary structure elements, the native structure of many small proteins is stabilized by the binding to metal ions or by the presence of disulfide bonds in close spatial vicinity (69). The covalent linkage of cysteine residues by disulfide bonds is an important and, in many cases, essential structural feature for numerous proteins (1, 3, 31, 58). Disulfide bonds usually cross-link distant regions of the protein sequence, decreasing the entropy of the unfolded state and making it less favorable relative to the folded conformation (7, 83). Sometimes, disulfide bonds also act enthalpically, by stabilizing local interactions (83). Although in most cases disulfides play a structural role and are, therefore, only indirectly essential for protein function, there are also examples in which they are involved in the regulation of protein activity (10, 48).

Small protein domains whose structural features are determined by their disulfide bonds are usually referred to as disulfide-rich domains (DRDs) (24). The group of proteins that is covered by this definition is broad, including both secreted and intracellular proteins, and they are involved in a wide variety of functions, such as growth factors, toxins, enzyme inhibitors, and structural or ligand-binding domains within larger polypeptides (24). Since protein classification of small proteins is often unreliable using common sequence and structural comparison tools, Cheek *et al.* undertook a classification of 2945 DRDs found in the PDB on the basis of their structural and evolutionary relatedness as well as disulfide bonding patterns. Their classification leads to the recognition of 41-fold types (24). Domains in the same fold group share a structural core composed of secondary structure elements found in the same spatial arrangement. In this work, we exploit this database to investigate the folding and aggregation properties of DRDs, providing new insights on how the interplay between these properties shapes their structures and influences their folding pathways.

Departament de Bioquimica i Biologia Molecular, Institut de Biotecnologia i Biomedicina, Universitat Autònoma de Barcelona, Barcelona, Spain.

*These two authors contributed equally to this work.

Innovation

DRDs are small versatile proteins involved in a wide range of functions. The structuration of these kinds of proteins depends on the formation of a specific pattern of disulfide bonds through oxidative folding reactions, which are extremely slow when compared with the folding kinetics of globular proteins. This implies that DRDs populate unfolded or partially folded ensembles during relatively long periods of time, which might favor side aggregation reactions, resulting in an extremely dangerous situation in the cell. We demonstrate, however, that DRDs sequences have evolved to support a particularly low intrinsic aggregation load and, in fact, their solubility properties resemble those of IDPs, which constitute a paradigm of protein solubility. Moreover, we provide a rationale to forecast and explain the oxidative folding pathways of DRDs, a goal that remained elusive to date, on the basis of the interplay between folding and aggregative properties.

Results

In the absence of disulfide bonds, DRDs are predicted to be intrinsically unstructured

The structure of a protein results from the balance between the conformational entropy and the energy of residue interactions (39). During protein folding, the formation of a sufficient number of interactions is necessary to compensate the loss of conformational entropy. In this context, the influence of covalent disulfide bonds to protein stability is unique, as they mainly act by decreasing the entropy of the unfolded polypeptide (66). This, however, raises the question of what is the structure adopted by DRDs in their reduced state, that is, in the absence of the entropic strain imposed by the disulfide bonds. In order to address this issue, we undertook a series of *in silico* analysis using the Cheek's dataset (24). In this dataset, the different DRDs are classified according to their structural features, enabling the selection of proteins covering the complete spectrum of structures adopted by this group of proteins. Within the fold groups, the domains are assembled into families of homologs. Accordingly, the complete dataset could be classified in 98 different families on the basis of structural and evolutionary relationships (24). We used a representative protein of each of these families for further analysis. The insulin-like family representative was, however, removed, as it consists of two different polypeptide chains, thus resulting in a dataset of 97 DRDs (Supplementary Table S1; Supplementary Data are available online at www.liebertpub.com/ars).

First, we analyzed the presence of disorder in the sequences of DRDs using the FoldUnfold server (42), assuming that the proteins are in the reduced state. This algorithm exploits the computed mean packing density in a globular state for each of the 20 natural amino-acid residues to generate packing density profiles and identifies disordered segments as those protein regions enriched in residues displaying low packing density values (42). We detail the results of such analysis in Supplementary Table S2. Interestingly, 41% of DRDs sequences are predicted to be unfolded in the reduced state by this algorithm (Fig. 1A). We mapped the predicted disordered regions over the secondary structure elements in the DRDs structure (Table 1). A majority (58%) of the predicted disordered residues map to unstructured regions in the native, disulfide-bonded conformation, another 28% to regular secondary structure elements, and the rest to turns.

Cys, Trp, Tyr, Ile, Phe, Val, and Leu are considered order-promoting residues, as they occur more frequently in ordered than in disordered structural regions (37, 72). In addition, in the FoldUnfold scale, Cys has a high mean packing density (42), reflecting that, in protein structures, Cys residues tend to exist in densely packed environments. However, in contrast to the rest of order-promoting residues, which are bulky and hydrophobic, in protein structures, Cys residues are closer to other Cys than to any other type of residue, including hydrophobic ones (59). This suggests that Cys are packed, because they can form disulfides and not due to any other intrinsic physicochemical property of this amino acid. Therefore, it is predicted that under reducing conditions, Cys would contribute less to protein order.

Despite there being no good substitution for Cys, it can tolerate substitutions with other small amino acids and in protein engineering studies, Cys is preferentially mutated to Ser or, alternatively, to Ala. These substitutions are considered

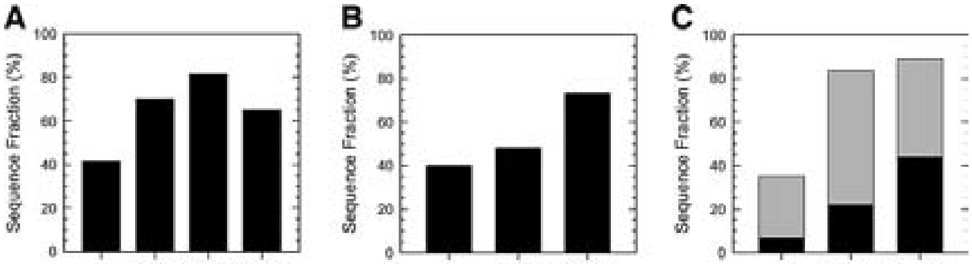


FIG. 1. Prediction of protein disorder in DRDs. Percentage of DRDs and their mutated variants predicted to be disordered according to **(A)** FoldUnfold, **(B)** FoldIndex, and **(C)** RONN algorithms. The fraction of DRDs sequences predicted to be unfolded is represented by black bars. In **(C)**, the remaining domains with at least 10 consecutive residues predicted as disordered are represented by gray bars. Disorder predictions are shown for wt domains, DRDs mutants with cysteine mutated to alanine (Cys/Ala), and variants with cysteine mutated to serine (Cys/Ser). In **(A)**, predictions for DRDs with cysteine mutated to a hypothetical amino acid X with the average packing density of the 20 natural amino acids (Cys/X) are also shown. wt, wild-type; DRD, disulfide-rich domain; RONN, regional order neural network.

Table 1. Distribution of Disordered Residues in DRDs, as Predicted by FoldUnfold, Among the Different Secondary Structure Elements in Their 3D Structures

SS type	*Ratio of residues (%)*
Helix	18.06
Strand	9.56
Turn	14.76
Unstructured	57.62

DRD, disulfide-rich domain; SS, secondary structure.

the less disruptive to study the role of covalent bonding in protein structures. We virtually mutated Cys to Ala and Ser in the 97 domains and re-analyzed the degree of predicted disorder (Supplementary Table S2). The 69% and 82% of the sequences in which Cys are mutated to Ala and Ser are now predicted to be unstructured, respectively (Fig. 1A). We also considered the mutation of Cys residues in DRDs to a hypothetic amino acid X displaying the average packing density of the 20 natural residues. In this case, 64% of the sequences are predicted to be unfolded (Fig. 1A).

In order to validate the results obtained with FoldUnfold we employed another two additional protein disorder detection tools: FoldIndex (70) and regional order neural network (RONN) (85), based on unrelated prediction models (Supplementary Tables S3 and S4).

The FoldIndex server developed by the Sussman lab (70) exploits an algorithm previously described by Uversky *et al.* (80) that takes into account the average hydrophobicity of natural amino acids in the Kyte-Dolittle scale and the average net charge of the polypeptide. FoldIndex provides a single score for the entire sequence, predicting whether it is folded (positive values) or not (negative values). Assuming that the proteins are reduced, it predicts 41% of DRDs sequences as intrinsically unstructured (Fig. 1B). When we used FoldIndex to analyze the effect of Cys to Ser mutations in the foldability of DRDs (Supplementary Table S3), 74% of the mutated sequences were predicted be disordered (Fig. 1B). When we considered Cys to Ala mutations, 49% of the proteins still displayed negative values (Fig. 1B), despite Ala being considered a hydrophobic, order-promoting amino acid, in the FoldIndex scale (70).

The RONN algorithm developed by Yang *et al.* (85) aligns a query protein sequence to a set of prototype disordered/ordered regions and uses the alignment scores to classify the query sequence. RONN predicts 7% of reduced DRDs to have all of their residues unstructured, and an additional 29% are predicted to contain unstructured segments comprising more than 10 consecutive residues (Fig. 1C). The 21% and 45% of the sequences in which Cys are mutated to Ala and Ser are predicted as totally unstructured, respectively (Fig. 1C). An additional 61% and 45% of the sequences contain disordered regions in Ala and Ser mutants, respectively (Fig. 1C).

Venn diagrams analysis of the predictions for wild-type domains indicates a ~60% overlap between methods when compared in pairs (Supplementary Fig. S1).

In order to confirm that the predicted large increase in disorder on Cys to Ser mutation does not arise simply from a higher number of Ser residues in the mutated sequences, an effect that could be potentially important in short proteins, we analyzed whether there was any correlation between the length-normalized increase in the % of residues predicted to be disordered on mutation and the number of mutated Cys in a sequence. Pearson correlation coefficients of $r^2=0.008$, $r^2=0.004$, and $r^2=0.007$ were obtained for FoldUnfold FoldIndex and RONN, respectively, enabling us to discard this effect.

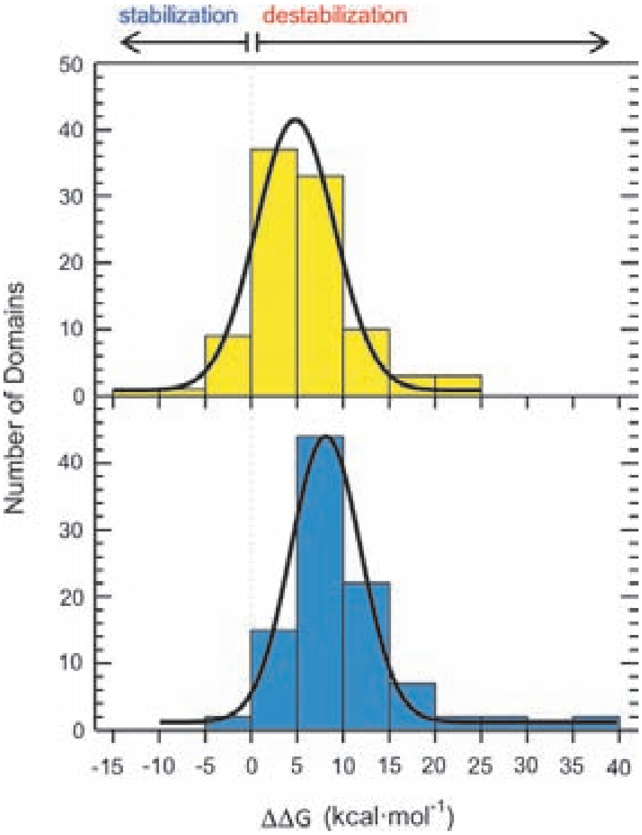


FIG. 2. Distribution of changes in DRDs stability upon mutation. Changes in free energy of unfolding (ΔΔG) computed using the FoldX forcefield for Cys to Ala (yellow) and Cys to Ser (blue) mutants. Stability changes are represented as histograms that were calculated using 5 kcal·mol^{-1} bins, and histogram data were fitted to Gaussian distributions (black lines). To see this illustration in color, the reader is referred to the web version of this article at www.liebertpub.com/ars

The analysis of 100 randomly selected sequences shorter than 150 residues and devoid of disulfide bonds from the SCOP-derived ASTRAL40 dataset (15) with FoldUnfold and FoldIndex indicates that, as expected, reduced DRDs and especially their Ala and Ser analogues are predicted to display a higher degree of disorder than these domains (Supplementary Fig. S2).

Overall, independently of the considered algorithm and mutational scheme, these data converge to indicate that, in the absence of disulfide bonds, many DRDs would resemble intrinsically disordered proteins (IDPs) in structural terms.

In the absence of disulfide bonds, DRDs are predicted to be highly destabilized

A reduction of disulfide bonds usually leads to DRDs global unfolding (18, 30, 53). We calculated the theoretical differential stability of analogues of the 97 domains, in which

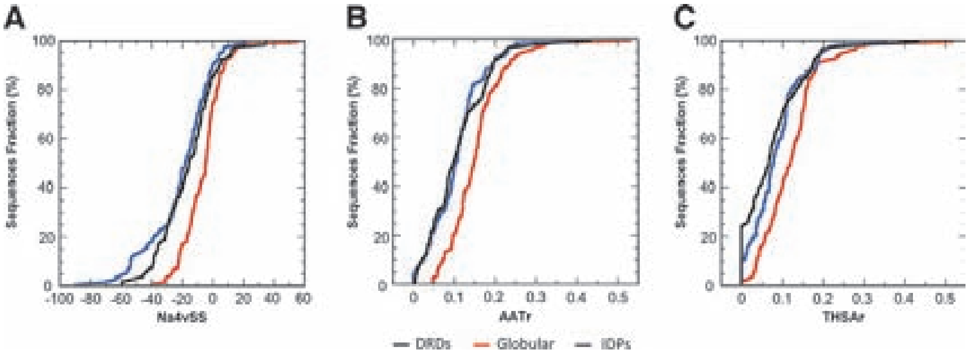


FIG. 3. Comparison of the aggregation properties of DRDs, small globular proteins, and IDPs. Value distribution of different parameters indicative of protein aggregation capability calculated using the AGGRESCAN algorithm: **(A)** average aggregation propensity (Na4vSS), **(B)** area of the aggregation profile above the threshold per residue (AATr), and **(C)** total area of the aggregation peaks – Hot Spots – per residue (THSAr). The distributions are shown in black for DRDs and in red or blue for the datasets of globular proteins or IDPs, respectively. IDP, intrinsically disordered protein. To see this illustration in color, the reader is referred to the web version of this article at www.liebertpub.com/ars

all the Cys residues involved in disulfide bonds (592 residues from a total of 613 Cys) were mutated to either Ala or Ser, using FoldX, a force field that enables a structure-based estimation of mutational free energy changes on the stability of a protein (44, 76). The histograms of $\Delta\Delta G$ values for the 194 resulting variants reveal a broad distribution with peaks around $+5$ and $+8\,\mathrm{kcal \cdot mol^{-1}}$ for Ala and Ser mutants, respectively, with a tail toward more destabilizing $\Delta\Delta G$ values, especially in the case of Ser variants (Fig. 2) and a main contribution of entropic factors (Supplementary Fig. S3). $\Delta\Delta G$ values $\geq 5\,\mathrm{kcal \cdot mol^{-1}}$ can be considered strongly destabilizing (78) and might well account for the disruption of DRDs structure on reduction. No significant correlation ($r^2 = 0.025$) was found between the predicted increase in disorder in the sequence on Cys to Ser mutation according to FoldIndex and the calculated FoldX destabilization, as the energetic contribution of native disulfides depends on their topological connectivity. In addition, we did not observe any obvious relationship between the calculated destabilization and the specific fold or function of the domain.

DRDs sequences display low aggregation propensity and reduced binding sites to Hsp70 chaperones

The failure of proteins to fold, or to remain correctly folded, usually results in their aggregation, which is associated with the impairment of essential cellular processes. Protein aggregates might elicit stress responses, inhibit the proteasome, sequestrate chaperones and transcription factors, promote membrane pore formation and calcium overload, and cause oxidative stress and mitochondrial dysfunction, among other deleterious effects, being thus associated to pathological states (25). Recent results show that protein aggregation can be reliably correlated to a set of defined physicochemical parameters (26) and determined to a large extent by the primary sequence (49, 82). This has enabled the development of algorithms that are able to predict the presence of aggregation-prone regions (APRs) in proteins as well as their overall aggregation propensities (14).

As a general trend, evolution has endorsed globular proteins with solubility in their functional conformations. Nevertheless, when their stability is compromised and they become partially unfolded, the exposition of previously protected APRs might trigger aberrant aggregation reactions (13, 38). Our results indicate that before they can form their disulfide bonds or in the eventual case that they become reduced, DRDs have both a high sequential and energetic propensity to occupy unfolded/disordered states, which implies that they might be at risk of aggregation. We used AGGRESCAN, an algorithm previously developed by our group, to predict protein aggregation in this protein set (11, 27, 34). We addressed how the aggregation properties of DRDs sequences can be compared with those of intrinsically disordered and globular proteins of a similar size. To this aim, we used the DISPROT (77) and the SCOP-derived ASTRAL40 datasets (15) and randomly selected 100 sequences shorter than 150 residues and devoid of disulfide bonds from each database. The following parameters were calculated for the three groups of proteins (all normalized by the protein length):

(i) The average aggregation propensity of the sequence (Na4vSS)
(ii) The number of aggregation-prone regions (NnHS)
(iii) The average aggregating potency of the detected aggregation peaks (THSAr)
(iv) The average aggregating potency of residues above the detection threshold (AATr), independently whether they are clustered in aggregating peaks or not.

DRDs display obviously lower average aggregation propensities than proteins from the ASTRAL40-derived dataset (Fig. 3A), with mean Na4vSS values of -16.4 and -6.3, respectively (the more negative the value, the more soluble the protein). A similar trend is observed for AATr, which is yet another measure of the overall aggregation propensity of the sequences (Fig. 3B). In addition, only 2% of the proteins from the ASTRAL40-derived dataset are devoid of aggregating stretches, whereas this value increases and includes 22% of the DRDs. In fact, 82% of the proteins devoid of APR were predicted as unfolded by FoldUnfold. In the same line, those DRDs presenting APRs have an average of 2.9 APRs for each 100 residues in contrast to the 3.9 APRs predicted for the ASTRAL40 group. Not only DRDs display fewer APRs, but also the average aggregating potency of the detected aggregation peaks in their sequences (THSAr) is lower (Fig. 3C). In fact, in terms of aggregation properties,

Table 2. Aggregation Properties of the IDPs, DRDs, and Globular Protein Datasets as Predicted by AGGRESCAN

	IDPs	*DRDs*	*Globular*
Na4vSS	−21.2	−16.4	−6.3
AATr	0.106	0.109	0.154
NnHS	2.557	2.950	3.942
THSAr	0.085	0.076	0.121

IDP, intrinsically disordered protein.

DRDs sequences resemble much more those of IDPs than those of globular proteins (Fig. 3A–C and Table 2).

To ascertain whether the reduced intrinsic aggregation properties of DRDs might result from an amino-acid compositional bias, we compared their amino-acid content with the average composition of the proteins in the Swiss-Prot database (Fig. 4). Following the trend described earlier, DRDs are significantly depleted in Ile, Leu, and Val, three aliphatic residues that display high-aggregation propensity in most aggregation scales. In globular proteins, these residues are usually buried inside the tertiary structure and their depletion in DRDs would respond to the absence of hydrophobic cores in these disulfide-bonded domains. This might contribute to the lower aggregation propensity of DRDs. In contrast, they are enriched in Pro, a residue incompatible with the formation of β-sheet-enriched aggregates.

Overall, the analysis of the aggregating properties of DRDs suggests that as previously reported for IDPs (54), their sequences might have been under evolutionary control to minimize the presence of aggregation-prone residues as a strategy to prevent the aggregation of the solvent-exposed polypeptide chain before oxidative folding results in the attainment of a globular protective conformation.

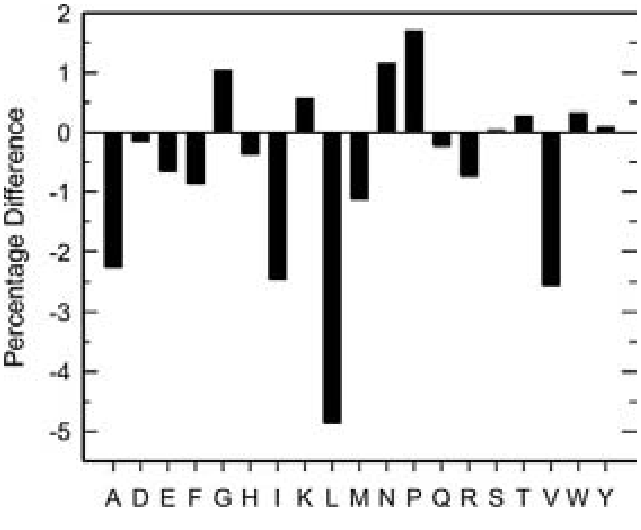


FIG. 4. Comparison of the amino-acid composition of DRDs and small globular proteins. Differential composition of DRDs relative to the Swiss-Prot database 2013. Differential abundances are shown for each natural amino acid except cysteine (included in the computational analysis but not represented).

Disordered proteins have been shown to require, in general, less assistance from chaperones than ordered, globular proteins (47). To explore whether this could be the case of DRDs, we used LIMBO, a chaperone binding site predictor for the Hsp70 chaperones, trained from peptide binding data and structural modeling (81). We analyzed potential Hsp70 binding sites in the IDPs, DRDs, and ASTRAL40-derived datasets. As can be seen in Figure 5, both IDPs and DRDs display a significantly reduced number of Hsp70 binding sites when compared with globular proteins, reinforcing the structural similitude between reduced DRDs and disordered polypeptides. Whether DRDs exhibit also low binding to members of chaperone families other than Hsp70 should be further investigated.

Amyloidogenic regions in DRDs overlap with secondary structure elements

Albeit aggregation-prone and amyloidogenic regions coincide in many cases (73), AGGRESCAN is not aimed at identifying short sequences that are able to nucleate the specific and ordered assembly of amyloids. This task is better attained using the WALTZ algorithm, as it employs a position-specific scoring matrix deduced from the analysis of the amyloid properties of a large set of hexapeptides (60). Using this program, we surveyed DRDs for the presence of potential amyloidogenic stretches. A total of 24 hits were obtained in 18 out of the 97 DRDs sequences. The same analysis in the ASTRAL40-derived dataset rendered 82 hits in 54 proteins, thus confirming the intrinsic low aggregation propensity displayed by DRDs.

Despite their low number, the detected sequences might potentially nucleate amyloid assembly if exposed to solvent. We explored the structures of the 18 DRDs and 54 globular proteins to map the detected amyloid stretches over their secondary structure elements. In both cases, the majority of residues in amyloid regions map to regular secondary structure elements (Table 3). This association between foldability and amyloidogenicity suggests a negative selection process acting to avoid the accumulation of aggregating sequences in unstructured regions. Interestingly enough, in DRDs, 44% of the amyloidogenic residues map to regions that adopt α-helical conformations in the native state and only 19% map to β-sheets, which is consistent with the view that one of the most effective strategies to neutralize amyloidogenic stretches in proteins is to hide them into α-helical structures (52, 79). This bias toward α-helices is not evident in the ASTRAL40 dataset, in agreement with previous data indicating that the aggregation propensity of all-α, all-β, and mixed α/β globular proteins is fairly similar (54).

Association between DRDs' predicted intrinsic properties and folding pathways

The folding and unfolding reactions of a large number of disulfide-containing proteins have been characterized (9, 58). These studies illustrated a high degree of diversity of oxidative folding pathways, especially for DRDs. The two extreme mechanisms are illustrated by (a) proteins that fold through a reduced number of intermediates containing mainly native disulfide bonds and (b) proteins that fold *via* a highly heterogeneous population of intermediates containing

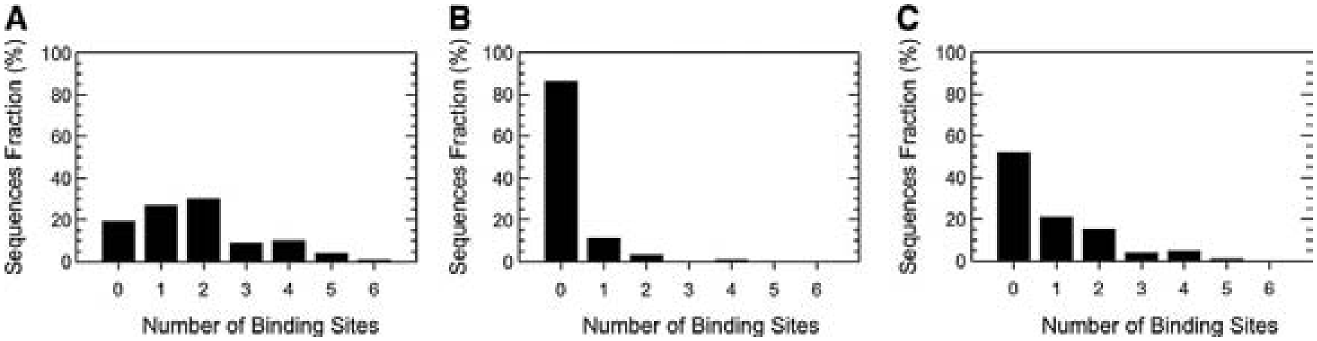


FIG. 5. Predicted chaperone binding to DRDs, small globular proteins, and IDPs. Distribution of the number of chaperone binding sites per protein predicted with LIMBO for **(A)** the dataset of globular proteins, **(B)** DRDs, and **(C)** the dataset of IDPs.

mostly non-native disulfides, including fully oxidized scrambled isomers (19, 20). Leech-derived trypsin inhibitor (LDTI) and hirudin (Fig. 6A, B) represent two well-characterized models folding through the (a) and (b) mechanisms, respectively (4, 22, 23, 67).

To see whether there exists any relationship between the folding mechanism of a DRD and its intrinsic properties, we analyzed the sequences of LDTI and hirudin using FoldIndex (Fig. 6A, B). Eighty-nine percent of the LDTI sequence is predicted to be folded (unfoldability = 0.20). In contrast, 71% of the hirudin sequence is predicted to be unfolded (unfoldability = − 0.14). The same trends are observed with FoldUnfold and RONN, which predict 13% and 0% of LDTI sequences to be disordered, respectively; however, for hirudin, these values increase to 32% and 45%, respectively. We used AGGRESCAN to compare the aggregation properties of the sequences of these two proteins. Reduced hirudin is predicted to be much more soluble than LDTI, with Na4vSS values of − 33.3 and − 3.2, respectively. No APRs are predicted for hirudin. In contrast, residues 23–27, overlapping the only α-helix in LDTI (residues 24–29), are predicted to be aggregation prone. In a previous work, we isolated and structurally characterized all the major intermediates in the LDTI folding pathway (67). Interestingly, all of them share a preformed α-helix, independently if it was covalently linked to rest of the structure or not, indicating that it forms early in the folding reaction.

The ligand binding module five (LA5) of the low-density lipoprotein (LDL) receptor is a DRD of 40 residues and three disulfide bonds with a calcium binding motif that is essential for its structure. We have shown that in the presence of calcium, such as hirudin, its folding involves an initial phase of non-specific packing that gives rise to complex equilibrated populations of intermediates (2). This results in an inefficient folding reaction that requires 36 h to be completed. More importantly, in the absence of calcium, native LA5 is not formed and the oxidative folding reaction ends up with the formation of non-native scrambled forms through a stochastic process, indicating that non-covalent interactions, if they occur, are not sufficient to drive the folding of this domain. Accordingly, FoldIndex predicts 52% of the sequence to be unfolded (Fig. 6C), with an unfoldability value of − 0.12. FoldUnfold and RONN predict 48% and 51% of LA5 to be disordered in the reduced form. AGGRESCAN predicts LA5 to be highly soluble with an Na4vSS value of − 36.1, and WALTZ does not detect any amyloidogenic stretch. Thus, for LA5 also, disorder and solubility appear to be associated with an inefficient folding pathway. The LDL receptor contains seven LA repeats in tandem comprising residues 25–313. The average Na4vSS value for this large region is − 32, with an unfoldability value of − 0.10 and absence of amyloidogenic sequences. These values might explain why in intact cells the individual modules of the LDL receptor do not fold independently and first collapse into folding intermediates that are characterized by long-distance non-native disulfide bonding and absence of native structure and why the receptor does not aggregate in this misfolded state (51).

Fuller and co-workers have recently compared the oxidative folding of four conotoxins, GIIIA, PIIIA, SmIIIA, and RIIIK, sharing identical disulfide connectivity (41a). After 120 min of a reaction, more than 60% of initially reduced RIIIK attained the native state. In contrast, less than 30% of the other proteins were folded at this time point, with a large accumulation of scrambled isomers. Again, we found a correlation between the efficiency of the reaction and the predicted aggregation and disorder properties of the proteins. In this way, the AGGRESCAN Na4vSS value of RIIIK is 4.3; whereas it is − 27.9, − 49.0, and − 52.7 for PIIIA, SmIIA, and GIIIA, respectively. Similarly, the FoldIndex unfoldability value of RIIIK is 0.21; whereas for the other proteins, this value ranges between − 0.37 and − 0.43.

Even with DRDs that share sequence homology and 3D structure, it has been so far impossible to forecast whether they would fold by similar or different pathways (9). This is best illustrated by the cases of bovine pancreatic trypsin

Table 3. Distribution of Proteins Amyloidogenic Residues in DRDs and Globular Proteins Datasets, as Predicted by WALTZ, Among the Different Secondary Structure Elements in Their 3D Structures

	Ratio of residues (%)	
SS type	*DRDs*	*Globular*
Helix	43.52	35.37
Strand	18.65	37.69
Turn	10.88	4.63
Unstructured	26.94	22.31

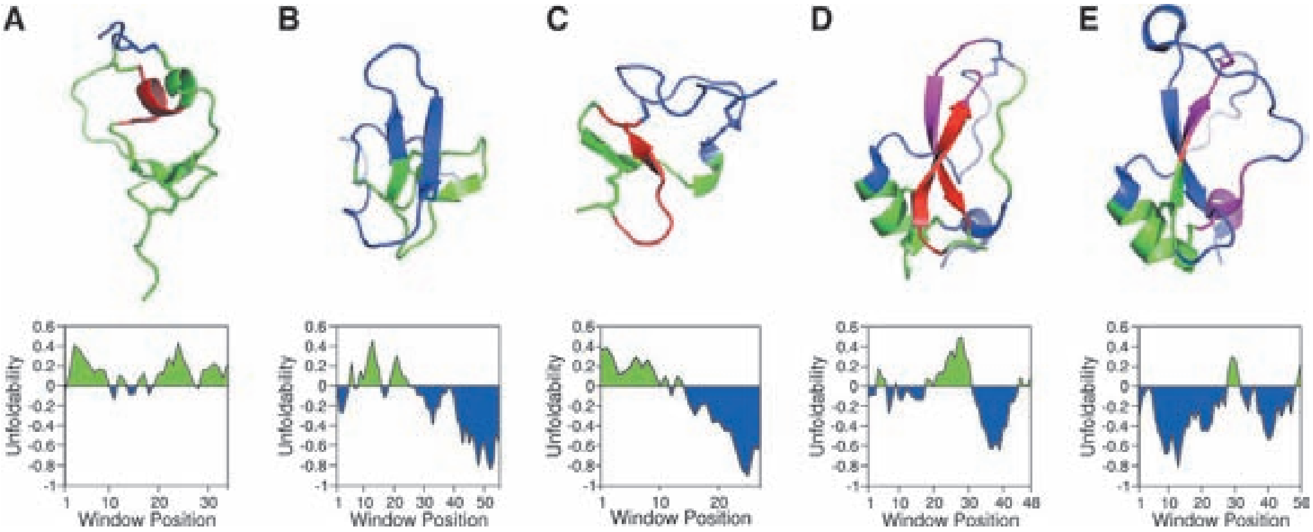


FIG. 6. Relationship between DRDs aggregation properties and their foldability. APRs detected by employing the AGGRESCAN algorithm (in red), disordered segments predicted with FoldIndex (in blue), and domain regions where APRs and disordered fragments overlap (in magenta) are mapped over **(A)** LDTI (PDB: 2kmo), **(B)** hirudin (PDB: 2hir), **(C)** The ligand binding module five (LA5) of the low-density lipoprotein receptor (PDB: 1AJJ), **(D)** BPTI (PDB: 1d0d) and **(E)** TAP (PDB: 1d0d) structures. Unfoldability profiles computed with FoldIndex are shown below its corresponding structure for each domain; green and blue indicate folded and disordered regions, respectively. APRs, aggregation-prone regions; BPTI, bovine pancreatic trypsin inhibitor; LDTI, leech-derived trypsin inhibitor; TAP, tick anticoagulant peptide. To see this illustration in color, the reader is referred to the web version of this article at www.liebertpub.com/ars

inhibitor (BPTI) and tick anticoagulant peptide (TAP), two proteins that share a very similar structure and identical disulfide connectivity (Fig. 6D, E) but exhibit different oxidative folding pathways (17, 21, 84). Folding of BPTI proceeds *via* an orderly formation of native 1- and 2-disulfide bonds and native-like subdomains. Out of 74 possible 1-, 2-disulfide isomers that may potentially serve as folding intermediates of BPTI, only 5 intermediates were shown to populate the pathway and all of them adopt native disulfide bonds. Fully oxidized 3-disulfide scrambled isomers are absent. In contrast, folding intermediates of TAP consist of a highly heterogeneous population of 1- and 2-disulfide isomers, with at least 30 species identified. More importantly, 3-disulfide isomers serve as major folding intermediates in TAP, the reshuffling of which is required to attain the native conformation. Despite TAP and BPTI structures exhibiting an RMS deviation of only 1.01 Å at the backbone level, the two proteins differ in the degree of predicted disorder. FoldIndex predicts 80% and 49% of TAP and BPTI residues to be disordered, with unfoldability values of −0.10 and −0.01, respectively (Fig. 6D, E). FoldUnfold does not predict any disordered region in the BPTI sequence, whereas 16% of the TAP sequence is disordered. RONN does not predict disordered regions in any of the two sequences. The two proteins are also expected to differ in their degree of flexibility in the oxidized state or extent of hydrogen–deuterium exchange, according to the H-protection server (56), which predicts 76% and 54% of the main chain hydrogen atoms to be exchangeable in TAP and BPTI, respectively; suggesting that, despite their structural similarity, the different degree of disorder encoded in their sequences might also impact the conformational properties of their folded states. To further analyze this issue, we used CABS-flex (50), an efficient procedure for the simulation of structure flexibility of folded globular proteins. The analysis of TAP and BPTI structures with CABS-flex confirms that TAP experiences larger conformational fluctuations than BPTI, a property that can be more obviously appreciated in the residue mean-square-fluctuation profile of the respective simulations (Supplementary Fig. S4).

AGGRESCAN predicts TAP sequence to be much more soluble than that of BPTI with Na4vSS values of −24.7 and −7.0 for TAP and BPTI, respectively. In BPTI, the regions encompassing residues 16–24 and 28–36, which overlap almost exactly with the two central β-sheets, are predicted to be aggregation prone (Fig. 6D). WALTZ also predicts the second stretch (residues 29–37) to be amyloidogenic. Interestingly, the two β-sheets in BPTI have been shown to constitute the nucleus from which conformational folding initiates (12). AGGRESCAN does not detect any APR in the first β-sheet of TAP. In addition, the aggregation-prone 35–39 region overlapping with the end of the second TAP β-sheet is predicted to display lower potency than the correspondent region in BPTI. WALTZ does not detect any amyloidogenic region in TAP. In order to analyze whether our predictions reflect experimental aggregation amyloid propensities, we synthesized two peptides corresponding to β-sheet 1 (18-IIRYFYN-24) (β1-BPTI) and β-sheet 2 (22-LCQTFVY-35) (β2-BPTI) in BPTI and two peptides corresponding to β-sheet 1 (22-ERAYFRN-28) (β1-TAP) and β-sheet 2 (32-GCDSFWI-38) (β2-TAP) in TAP and analyzed their *in vitro* aggregation properties.

β1-BPTI and β1-TAP peptides were incubated at 100 μM at pH 7.5 and 25°C for 48 h, and their aggregation was evaluated using synchronous light scattering (Fig. 7A). In excellent agreement with the predictions, β1-BPTI exhibited two times higher scattering signals than β1-TAP. We also monitored the binding of the two incubated peptides to the

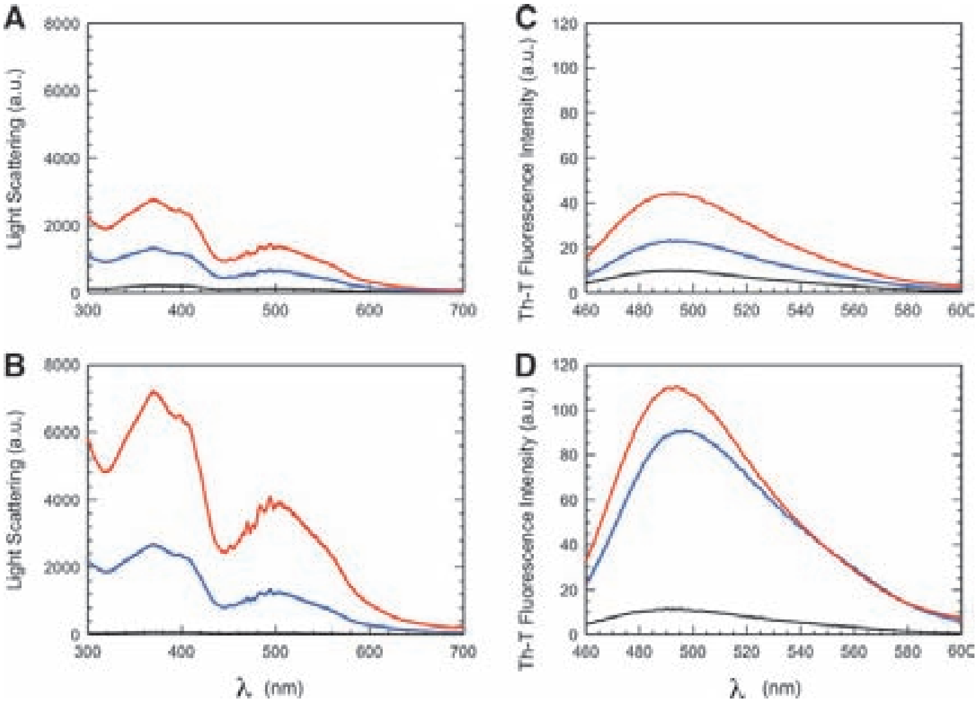


FIG. 7. Light scattering and Th-T binding of BPTI and TAP β-sheet peptides. Light scattering after incubation of **(A)** *β*1 peptides from BPTI (red) and TAP (blue) and **(B)** *β*2 peptides from BPTI (red) and TAP (blue) in the presence of TCEP. Th-T binding after incubation of **(C)** *β*1 peptides from BPTI (red) and TAP (blue) and **(D)** *β*2 peptides from BPTI (red) and TAP (blue) in the presence of TCEP. Black lines represent buffer scattering and free Th-T fluorescence, respectively. TCEP, *tris*(2-carboxyethyl)phosphine; Th-T, Thioflavin-T. To see this illustration in color, the reader is referred to the web version of this article at www.liebertpub.com/ars

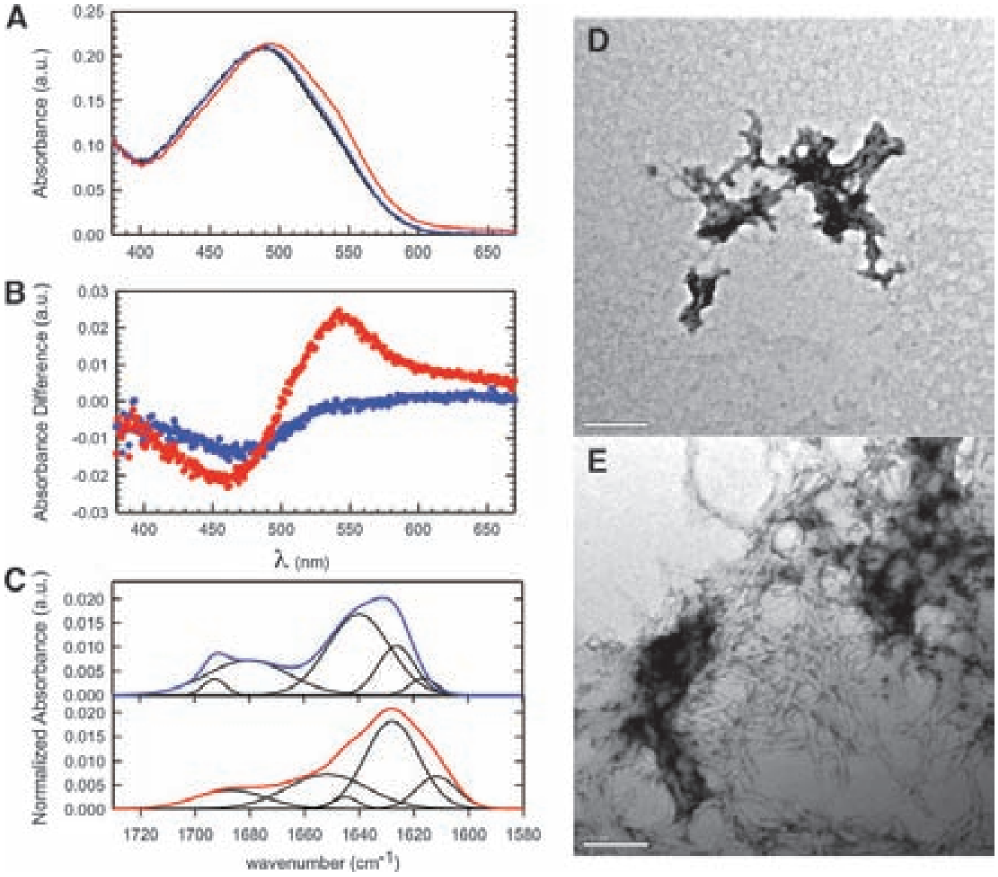


FIG. 8. CR binding, structural and morphological properties of BPTI and TAP β2 peptides incubated in the presence of TCEP. (A) Absorbance spectra of free CR (black) and after addition of *β*2-BPTI (red) or *β*2-TAP (blue) peptides after incubation. **(B)** Differential spectra of CR bound to *β*2-BPTI (red) or *β*2-TAP (blue) subtracted with free CR spectrum. **(C)** ATR-FTIR spectra of aggregated *β*2-BPTI (red) and *β*2-TAP (blue). IR spectra were deconvoluted to a maximum of five Gaussians. TEM micrographs of **(D)** *β*2-TAP and **(E)** *β*2-BPTI aggregates. CR, Congo red; ATR-FTIR, attenuated total reflectance–Fourier transform infrared spectroscopy; TEM, transmission electron microscopy. To see this illustration in color, the reader is referred to the web version of this article at www.liebertpub.com/ars

amyloid dye Thioflavin-T (Th-T) (Fig. 7C). Again, the increase in Th-T fluorescence emission elicited by the BPTI peptide was twofold that of the TAP peptide. However, none of the peptides promoted significant spectral changes in Congo red (CR), which is yet another amyloid dye (data not shown).

Since β2-BPTI and β2-TAP peptides contain a Cys residue, they were incubated as β-sheet 1 peptides, but 10 m*M* *tris*(2-carboxyethyl)phosphine (TCEP) was added in order to ensure their reduction. In agreement with the predictions, β-sheet 2 peptides exhibited significantly higher scattering signal and binding to Th-T than β-sheet 1 peptides (Fig. 7B, D). Again, the scattering and Th-T signals of β2-BPTI were higher than those of β2-TAP. Moreover, only the β2-BPTI peptide was able to promote a red shift in the CR spectra; the difference between spectra of the dye in the presence and absence of the peptide exhibiting the characteristic amyloid band at 540 nm (Fig. 8A, B). Attenuated total reflectance–Fourier transform infrared spectroscopy (ATR-FTIR) enabled addressing the structural features of the aggregated β-sheet 2 peptides. Decovolution of the absorbance spectra in the amide I region demonstrates significant differences in the secondary structure content of these aggregates (Fig. 8C). In this way, the IR spectrum of the β2-TAP is dominated by a signal at 1641 cm^{-1}, accounting for 48% of the spectrum area, typically attributed to unordered structures; whereas the IR spectrum β2-BPTI exhibits a mean band at 1628 cm^{-1}, which accounts for 45% of the total area, corresponding to a β-sheet secondary structure. The morphological features of incubated β-sheet 2 peptides were analyzed using transmission electron microscopy (TEM). As shown in Figure 8D and E, in both cases, we detected the presence of protein aggregates. Nevertheless, in good agreement with CR, IR data, and WALTZ predictions, the aggregates formed by β2-TAP were amorphous whereas those of β2-BPTI displayed an amyloid-like morphology. Overall, the experimental data confirm that the central elements of TAP structure are less aggregation prone than the topologically equivalent structures in BPTI.

Human proinsulin is a single polypeptide consisting of chains A and B which are connected by three disulfides and linked by a 35-residue-long connecting peptide that will be later cleaved off to yield mature insulin. The folding pathway of proinsulin involves a reduced number of intermediates and is, therefore, fast and efficient, being directed by the collapse of A and B chains to form a folding nucleus, with a negligible impact of the connecting peptide (46). This mechanism is correlated with the different predicted aggregation and disorder properties of the proinsulin domains. Chains A and B are predicted to be ordered by the three predictors and aggregation prone by AGGRESCAN, with unfoldability values of 0.34 and 0.21 and Na4vSS values of 15.9 and 10.9, respectively; whereas the connecting peptide is predicted to be disordered and highly soluble (Na4vSS = −34.7). Insulin-like growth factor-1 (IGF-1) shares proinsulin topology with two A and B chains that are linked by three conserved disulfide bonds with identical connectivity and a connecting peptide of 12 residues. In contrast to proinsulin, the folding of IGF-1 renders two different isomers, the native and a swapped form (46). Proinsulin and IGF-1 share most similarities structurally and functionally. Considering their high sequence homology and small size, their different folding behaviors are intriguing. While the predicted aggregation and disorder properties of the IGF-1 A chain and the connecting peptide are similar to those of proinsulin, the B chain is predicted to be less ordered and specially more soluble, with unfoldability and Na4vSS values of 0.12 and −1.4, respectively. These differences might well account for the different folding properties of these structurally similar proteins. A set of Proinsulin/IGF-1 hybrids has been constructed by exchanging their A and B chains, and their folding reactions are evaluated. According to our predictions, all molecules containing the less aggregation-prone B chain of IGF-1 folded into two isomers, whereas those containing the B chain of proinsulin folded into a unique stable structure (46).

Association between predicted intrinsic properties and the folding pathway of Nerita versicolor *carboxypeptidase inhibitor*

The data in the previous section suggest that DRDs, such as hirudin, LA5, or TAP, displaying sequences with significant disorder and low aggregation tendency might find it difficult to follow preferential folding pathways along which the protein folds with a high probability guided by the rapid formation of native-like local interactions. In these proteins, the folding reaction would occur through an initial stage of non-specific disulfide-bond bonding (packing), leading to the formation of scrambled species, which are further resolved into the native connectivity in an extremely slow reaction involving the establishment of non-covalent native-like contacts (3). In contrast, proteins such as LDTI and BPTI displaying lower disorder propensity and harboring sequences in which native secondary structure and aggregating regions overlap significantly would tend to fold through faster pathways that are dominated from the very beginning by non-covalent native-like interactions, resulting in the formation of a reduced number of folding intermediates in which native disulfides are predominant (3).

We explored whether this apparent association between predicted intrinsic protein properties and folding pathways would enable us to predict the way in which a previously uncharacterized DRD would fold. To this aim, we used *Nerita versicolor* inhibitor (NvCI), a novel 53-residue-long exogenous proteinaceous inhibitor of metallocarboxypeptidases from the marine snail *Nerita versicolor* (29). The NvCI structure defines a distinctive protein fold that is composed of a two-stranded antiparallel β-sheet connected by three loops (Fig. 9A). FoldIndex provides an unfoldability value of 0.10 for NvCI, FoldUnfold predicts the protein to be fully folded, and RONN predicts 85% of the sequence to be ordered. AGGRESCAN provides an Na4vSS value of −9.5 and predicts the regions encompassing residues 23–28 and 40–44, which overlap with the two central β-sheets, to be aggregation prone (Fig. 9A, B). WALTZ also predicts the second β-sheet (40-HVWTFE-45) to be amyloidogenic. Overall, the intrinsic properties of NvCI resemble much more those of LDTI/BPTI than those of hirudin/TAP proteins. To analyze whether, as predicted, this similitude results in an efficient folding reaction, we expressed, purified, and analyzed the conformational properties and oxidative folding of NvCI.

According to the crystal structure of NvCI (29), residues 41-VWT-43 in the amyloidogenic stretch are totally protected from the solvent in the native state, which would prevent the aggregation of the protein once it folds. We took profit of the buried Trp42 being the only tryptophan residue

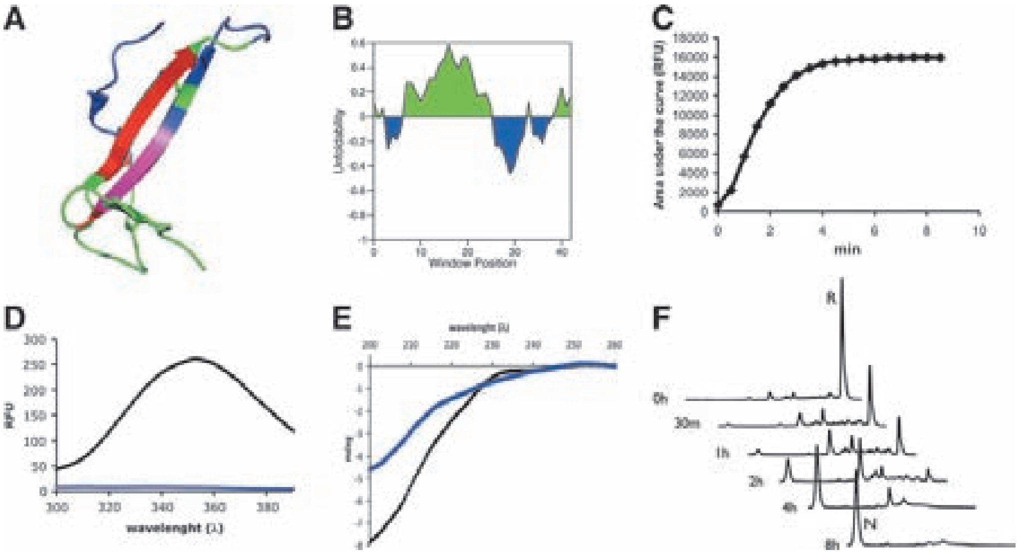


FIG. 9. Conformational properties and oxidative folding of NvCI. (A) APRs detected employing the AGGRESCAN algorithm (in red), disordered segments predicted with FoldIndex (in blue) and domain regions where APRs and disordered fragments overlap (in magenta) are mapped over the NvCI structure (PDB: 4A94). **(B)** Unfoldability profile computed with FoldIndex; green and blue indicate folded and disordered regions, respectively. **(C)** NvCI reductive unfolding was followed after TCEP addition. The normalized area under each Trp fluorescence spectrum was plotted as a function of time. **(D)** TCEP 20 m*M* was added to oxidized NvCI, and Trp emission spectrum was recorded demonstrating a large structural shift on protein reduction (native and reduced fluorescence spectra are shown in blue and black, respectively). **(E)** In agreement with Trp intrinsic fluorescence, significant changes in NvCI secondary structure are observed when oxidized NvCI is reduced with TCEP (native and reduced CD spectra are shown in blue and black, respectively). **(F)** Time-dependent NvCI oxidation was followed by acid trapping and RP-HPLC. As predicted from its intrinsic properties, NvCI displays fast kinetics and its oxidative folding pathway involves the formation of only two major intermediates. R and N indicate the reduced and native states, respectively. NvCI, *Nerita versicolor* inhibitor. To see this illustration in color, the reader is referred to the web version of this article at www.liebertpub.com/ars

in NvCI to monitor whether this region becomes exposed and, therefore, at risk of aggregation in the reduced state. To this aim, we added 20 m*M* TCEP to reduce the protein and monitored the change in Trp fluorescence along time. As can be seen in Figure 9C, an exponential increase in fluorescence emission occurs on reduction. A comparison of the Trp42 spectra in the oxidized and reduced states confirms that the initially buried residue becomes fully exposed once the protein loses its disulfide bonds (Fig. 9D). This transition is accompanied by structural changes leading to unfolded conformations, as evidenced in the circular dichroism spectra of the oxidized and reduced NvCI forms (Fig. 9E).

The oxidative folding of NvCI was examined with acid trapping and RP-HPLC analysis of the disulfide intermediates that accumulate along the folding reaction. The reduced and unfolded protein was allowed to refold in Tris-HCl buffer at pH 8.4 (see Materials and Methods section). As shown in Figure 9F, only a few major intermediates populate the folding process of NvCI, resulting in a fast and efficient reaction that is almost completed after 4 h, even in the absence of any thiol catalyst or oxidizing agent. Mass spectrometry analysis of the conversion of the major folding intermediate, containing two disulphide bonds, into native NvCI indicates that this step of the oxidative folding reaction is very fast and efficient (Supplementary Fig. S5), enabling a significant population of scrambled isomers to be discarded, which would trap the reaction as in the cases of hirudin, TAP, or the LA5 module. Thus, as predicted, the folding pathway of this DRD resembles in kinetic terms that of BPTI/LDTI-like proteins, supporting an association between DRDs intrinsic properties and oxidative folding pathways.

Discussion

Disulfide bond containing proteins are numerous and are involved in important biochemical processes. Extensive studies on the folding and conformational properties of selected DRDs have been performed *in vitro* under different experimental conditions (3). Here, we use a set of *in silico* approaches to systematically address the impact of disulfide bonds in these DRD features and experimentally validate theoretical predictions on selected DRD models. Oxidative folding catalysis occurs almost exclusively in the endoplasmic reticulum (ER), in the intermembrane space of mitochondria (41) in eukaryotes, and in the periplasm in bacteria (36). Systematic studies of DRDs folding rates *in vivo* are missing, mainly because disulfide bond formation occurs during translation and translocation, making the process complex to analyze. However, several data in the literature point out that under physiological conditions, DRDs oxidative folding is a relatively slow process. For mitochondrial DRDs, such as Cox19, oxidative folding and import are linked; the residence time of this protein in the cytosol in the reduced state has been shown to be 8 min, indicating that its *in vivo* oxidative folding rate is exceedingly slow (40). In the case of the LDL receptor, which contains several DRDs, it has been shown that, in the ER, several of these modules contain wrongly paired disulfides and require assisted reshuffling, which makes their folding slow (51). For BPTI, it has been shown that the effect of protein disulfide isomerase on the reduced protein and the 1-disulfide intermediates is negligible, suggesting that *in vivo* the first steps of oxidative folding are uncatalyzed and, therefore, slow (32). One of the best *in vitro* reconstitutions of a DRD *in vivo* folding reaction has been recently done by Purcell *et al.*, who evaluated the role of a number of ER-resident enzymes in the Conus venom glands in the folding of conotoxins produced by the same organism (74). The most favorable combination promoted folding of less than 40% of the total protein in 4 min. This rate and efficiency is very low when compared with those of globular proteins of the same size, which usually fold in the microseconds time scale (55). Thus, modeling DRDs properties in their reduced states is physiologically relevant, as it is likely that they populate this state significantly during the gap between ribosomal synthesis and the attainment of the native conformation at specific cellular compartments.

The comparative analysis of protein disorder and stability in wild-type and reduced DRDs analogues clearly indicates that, independently of the fold they adopt in the native state, most of these proteins would remain significantly unfolded before they can form the stabilizing covalent bonds. The persistence of such unfolded protein conformations is puzzling, as folding is known to be a primary strategy to prevent aggregation and, in globular proteins, even partial or transient unfolding might lead to pathological protein deposition (62). In DRDs, this danger is minimized, because they possess sequences with strongly reduced aggregation properties, relative to the remaining globular proteins. Experimental evidence supports the solubility of DRDs sequences. We had previously characterized the *in vitro* folding pathways of different DRDs (5, 6, 8, 67), with several of them involving the population of hundreds of transient, but long-lived, and unstructured intermediates. Despite refolding reactions from the initially reduced states being usually performed at high protein concentrations ($>1\,\mathrm{mg\cdot ml^{-1}}$) and some of them taking days to be completed in the absence of catalysts, we never observed side aggregation reactions. This behavior contrasts with the observation that accumulation of folding intermediates in globular proteins, even for small and fast folding polypeptides such as SH3 or WW domains, recurrently results in their aggregation (38, 45, 64).

In globular proteins, disulfide bonds strongly stabilize the structure, thus minimizing unfolding reactions that might expose previously hidden APRs (43). This is likely the underlying reason explaining why, as a trend, globular proteins with disulfides bonds tend to display more aggregation-prone sequences than proteins devoid of this bond (63). Our results indicate that this generic strategy is not enough, or is not acting, to ensure the solubility of DRDs, as their sequences seem to be shaped to exhibit low aggregation propensity and a reduced number of APRs despite being enriched in disulfides and highly stable in the oxidized state. This suggests that the low aggregation tendency detected in most of these domains is not intended to prevent aggregation on local unfolding of the native structure, but rather to maintain solubility during the process of oxidative folding. In fact, the presence of aggregating sequences in DRDs might have important consequences for their production. In this way, Hepcidin, a DRD hormone involved in iron homeostasis, displaying a strikingly high Na4vSS value of 17.1 is produced exclusively in inclusion bodies in bacteria and is resistant to solubilization on chemical synthesis (57). In yet another example, the aggregation during the recombinant production of eight Kunitz proteins, all sharing the same cysteine pattern, could be rationalized just using the AGGRESCAN output (75).

Once conscious that, in the reduced state, DRDs structurally resemble IDPs, the fact that their sequences tend to be soluble is not surprising. IDPs are known to display a high intrinsic solubility and a reduced number of APRs as a strategy to prevent the aggregation of the fully solvent-exposed polypeptide chain in the absence of a protective secondary structure (35, 54). The depletion in hydrophobic residues also prevents IDPs from chaperone-mediated protein degradation (65). All these properties seem to be shared by DRDs in their reduced states.

It is now clear that the ability to assemble into highly organized and cytotoxic amyloid-like structures is a feature shared by an extensive number of polypeptides (25). This implies that, apart from the native structure, an alternative, ordered, and stable state exists that may be accessible to proteins. These two states compete in the cell, because the set of non-covalent interactions that stabilize the native fold are also responsible for the formation of aggregates and, therefore, the formation of highly structured globular proteins comes at the expense of an increased aggregation propensity (54). Despite the DRDs in our dataset displaying a low content of amyloidogenic sequences and not being known to be involved in depositional disorders, the analysis of mature insulin, amylin, and calcitonin, three small proteins containing disulfide bonds involved in injection-localized amyloidosis, type 2 diabetes, and medullary carcinoma of the thyroid, respectively, indicate that, on the average, they display a low degree of predicted disorder along with high aggregation and amyloidogenic propensities (Supplementary Table S5), which suggests that these properties might also result in the formation of toxic assemblies when they converge in DRDs. The association between foldability and aggregation is evident in DRDs, where $\sim$62% of amyloid nucleating residues are located in sequences that adopt a regular secondary structure in the native, cross-linked, state. Interestingly, in DRDs, most of these regions reside in α-helices, likely because the presence of preformed β-sheets with high aggregation propensities is intrinsically dangerous, as shown for several disease-linked proteins (33) and here for the second β-sheet in BPTI. The bias toward the protection of amyloid regions in α-helices is apparently higher in DRDs than in globular proteins. Since the folding of α-helices essentially depends on local contacts, they are known to form faster than β-sheets; the adoption of helical structures early in the rather slow DRD oxidative folding reactions may provide a means to decrease the aggregation propensity of these amyloid stretches.

The oxidative folding pathways of the analyzed DRDs illustrate an association between their foldability and aggregation propensity. DRDs displaying high disorder propensity in their reduced state are intrinsically highly soluble and devoid of regions with high aggregation propensity, but this comes at the cost of inefficient folding pathways. On the contrary, the presence of APRs overlapping with structural elements in DRDs seems to be associated with faster and more efficient folding pathways in which preferential intermediates are sequentially formed, likely because hydrophobic APRs would contribute to form the folding nucleus from which the reaction extends. Our analysis suggests that in

DRDs, folding and aggregation properties are finely tuned. On the one hand, proteins displaying slow folding reactions should be intrinsically soluble, as otherwise the persistence of unfolded or partially folded intermediates will result in a potentially harmful aggregation. On the other hand, since efficient DRDs folding reactions seem to depend on the presence of hydrophobic stretches, they have been associated with a higher aggregative risk; thus, the folding reaction should be inherently efficient, in order to avoid the accumulation of dangerous intermediates and APRs should be protected preferentially within regular structural elements.

Materials and Methods

DRD dataset

All the analyses described next were performed over the sequences and structures of a set of 97 domain representatives of the DRDs families in which the 41 DRD folds are divided, according to the classification scheme developed by Cheek *et al.* (24). In addition, protein disorder and aggregation properties were analyzed for LDTI, hirudin, the LDL receptor module LA5, BPTI, TAP, and NvCI wild-type sequences.

Analysis of protein disorder in DRD sequences

Intrinsic disorder in DRDs was predicted using the web-based algorithms FoldUnfold (42), FoldIndex (70), and RONN (85) with default settings for both wild-type and mutant sequences. An 11-residue sliding window was used in all cases. Predictions were performed with the wild-type sequences of DRDs and with mutant sequences where Cys were substituted by Ala or Ser. For the analysis made with FoldUnfold, mutant DRD sequences were also generated where Cys were substituted by a hypothetic amino acid X that possesses the average packing density of the 20 naturally occurring amino acids (42).

Analysis of the thermodynamic impact of disulfide deletion in DRDs

The effect of disulfide bond disruption on DRDs thermodynamic stability was evaluated using the FoldX algorithm (76). DRDs mutants, where disulfide-forming Cys were substituted by either Ala or Ser, were generated with the BuildModel command. This command computes energetic changes on mutation by comparing the mutant's rotamer energetics relative to that of the wild-type structure. FoldX calculations were obtained as the average of five independent runs performed with default physicochemical and energetic parameters. Disulfide deletion effect on energetic parameters was computed using isolated DRDs structures retrieved from the PDB; in those cases where DRDs were found as a part of a multi-domain protein structure, DRDs were extracted by employing the domain boundaries defined by Cheek *et al.* (24).

Analysis of the intrinsic aggregation properties of DRDs

Different parameters indicative of protein intrinsic aggregation properties were predicted for DRDs using the AGGRESCAN algorithm web server (27) with default settings. Directions about the use of AGGRESCAN and its outcome interpretation may be found in the AGGRESCAN published tutorial (34). In order to compare DRDs aggregation properties with those of the globular and IDP ensembles, aggregation parameters were also calculated for two reference sets of 100 sequences with sizes below 150 residues and containing 1 or 0 Cys residues, which were retrieved among those proteins with less than 40% identity in the ASTRAL Compendium (ASTRAL40) (16) or among proteins in DISPROT database using pseudo-random numbers generated by the Mersenne Twister algorithm.

Identification of amyloidogenic regions in DRDs

The presence of amyloidogenic regions in DRDs and the ASTRAL40-derived dataset was evaluated using the web-based Waltz algorithm (60), employing default parameters.

Analysis of conformational flexibility in DRDs

The structural flexibility of DRDs was analyzed using the H-protection server (56) and the CABS-flex simulation algorithm (50), employing their default parameters.

Prediction of chaperone binding sites

Chaperone binding sites were predicted using the LIMBO algorithm webserver (81) employing default settings. Predictions were performed for the DRDs dataset and the sets of globular proteins and IDPs mentioned earlier.

Aggregation assay of BPTI and TAP derived peptides

Heptapeptides derived from BPTI and TAP corresponding to the sequences of strands β1 (β1-BPTI: IIRYFYN and β1-TAP: ERAYFRN) and β2 (β2-BPTI: LCQTFVY and β2-TAP: GCDSFWI) were purchased from EZBiolab, Inc. with a purity of >95%. Lyophilized peptides were dissolved in a 1:1 mixture of 1,1,1,3,3,3-hexafluoro-2-propanol (HFIP) and trifluoroacetic acid (TFA) in order to obtain 5 m*M* peptide solutions. Dissolved peptides were filtered through 0.22 μm filters to avoid the presence of pre-aggregated species and aliquoted to prepare peptide stocks. The solvent mixture was removed by centrifugal evaporation, and solid stocks were stored at −80°C. For aggregation assays, peptides were resuspended in 50 m*M* Tris pH 7.5, 150 m*M* NaCl to a 100 μM concentration and sonicated on an ice bath for 10 min before incubation. Peptide concentration was verified by absorbance spectroscopy at 280 nm, and peptide samples were allowed to aggregate at 25°C and 500 rpm agitation. After 48 h of incubation, peptide aggregates properties were evaluated. For β2 peptides, aggregation assays were performed in the presence of 10 m*M* TCEP to prevent the formation of intermolecular disulfide bonds.

Light scattering determination

Light scattering of incubated peptide samples was recorded from 300 to 700 nm at 90° using a FP-8200 spectrofluorimeter (Jasco) in synchronous mode. Scattered light spectra were obtained as the average of three consecutive scans.

Thioflavin-T binding

The binding of Th-T to incubated peptides was determined by diluting peptide samples in Th-T and recording the fluorescent emission spectra of the mixture between 460 and

600 nm, after equilibration at 298 K and employing an excitation wavelength of 445 nm in a FP-8200 spectrofluorimeter (Jasco). Final peptide and dye concentrations were 50 and 25 μM, respectively; and each spectrum was acquired as the accumulation of three consecutive scans.

Congo red binding

The binding of CR to aggregated peptides was evaluated by diluting fivefold the incubated peptides in CR, resulting in a final dye concentration of 10 μM, and recording the absorbance spectra of the mixture in the 380 to 670 nm range in a Cary 400 spectrophotometer (Varian).

Attenuated total reflectance–Fourier transform infrared spectroscopy

ATR-FTIR spectra were recorded for aggregated peptides samples as the accumulation of 16 scans in a Tensor 27 FTIR spectrophotometer (Bruker) using ATR mode default settings, and after evaporating water molecules under a N_2(g) stream. Infrared spectra were deconvoluted into overlapping Gaussian curves by employing the non-linear fitting program PeakFit (Systat Software).

Transmission electron microscopy

The morphology of the aggregated peptides was evaluated by TEM. Peptide samples were diluted 10-fold in milliQ water and then negatively stained by first allowing the deposition of 10 μl of the dilutions on carbon-coated copper grids for 5 min. After sample removal, the samples were stained for 1 min with 10 μl of 2% uranyl acetate, which was removed afterward. The grids were imaged at a 120 kV accelerating voltage in a JEM-1400 transmission electron microscope (JEOL).

Heterologous expression and purification of recombinant NvCI

The cDNA sequence of NvCI was fused in-frame to the *Saccharomyces cerevisiae* prepro-α-factor signal in the *Xho*I site of the pPICZαA vector for secretion into the culture medium. Production of recombinant NvCI was carried out using a *Pichia pastoris* Zeocin hyper-resistant strain in an autoclavable bioreactor (Applikon Biotechnology). Purification of NvCI was performed using a combination of two ion exchange chromatographic methods: an initial weak cation exchange (Accell™ Plus CM; Waters) using 20 m*M* Tris-HCl (pH 7.0) and an ionic strength gradient (up to 1 *M* NaCl), followed by a second step of anion exchange (TSKgel® DEAE-5PW; Tosoh Bioscience LLC) using a linear gradient of 0%–100% 20 m*M* Tris-HCl (pH 8.5) containing 1 *M* NaCl. The purity of NvCI was determined by its molecular mass obtained by MALDI-TOF-MS, by Tris/Tricine/sodium dodecyl sulfate polyacrylamide gel electrophoresis, and by its functional activity against bovine carboxypeptidase 1.

Oxidative folding

NvCI (0.3 mg·ml^{-1}) was reduced in 50 m*M* Tris pH 8.4, NaCl 100, and 50 m*M* DL-dithiothreitol (DTT) for at least 2 h at 23°C. To initiate folding, samples were passed through a PD-10 column (Sephadex-25; GE Healthcare) previously equilibrated with 50 m*M* Tris pH 8.4, NaCl 100 m*M*. To monitor the folding reaction, aliquots were removed at defined time points, quenched with 2% TFA, and analyzed by RP-HPLC as follows. Samples were injected in a Waters 2690 HPLC coupled to a UV detector set to 280 nm. A linear gradient of 20%–40% of 0.1% TFA in acetonitrile was applied for 70 min into a 250 × 4.6 (5μM) C4 column (Phenomenex) at a flow rate of 0.75 ml·m^{-1}. To monitor conversion between species, 1 μl of sample extracted at selected times from the oxidation reaction were derivatizated with 0.5 μl of 0.3 *M* vinylpiridine (Sigma) for 45 min, at room temperature, in darkness, then diluted to 15 μl with 0.1% TFA, and analyzed by mass spectrometry in a MALDI-TOF UltraFlextreme (Bruker Daltonics) in reflectron mode under 25 kV using external calibrators.

Intrinsic fluorescence

NvCI intrinsic tryptophan fluorescence was measured using NvCI (0.02 mg·ml^{-1}) after ON incubation with 20 m*M* TCEP (reduced state) and without any treatment (oxidized state). The spectra were measured in the 290–400 nm interval using a 280 nm excitation wavelength (10 nm excitation and emission slits, 0.1 s averaging time). NvCI reductive unfolding was analyzed by using tryptophan emission spectra every 30 s, as described earlier, after the addition of 20 m*M* TCEP.

Circular dichroism spectroscopy

Far UV CD spectra were acquired on a Jasco-710 spectropolarimeter that was continuously purged with nitrogen and thermostated at 25°C. Since both TECP and DTT interfered with far UV CD spectra, protein was reduced and before CD analysis, the reducing agent was removed using the protocol described earlier.

Acknowledgments

This work was supported by grants BFU2010-14901 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).

Author Disclosure Statement

No competing financial interests exist.

References

1. Anfinsen CB, Haber E, Sela M, and White FHJ. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain *Proc Natl Acad Sci U S A* 47: 1309–1314, 1961.
2. Arias-Moreno X, Arolas JL, Aviles FX, Sancho J, and Ventura S. Scrambled isomers as key intermediates in the oxidative folding of ligand binding module 5 of the low density lipoprotein receptor. *J Biol Chem* 283: 13627–13637, 2008.
3. Arolas JL, Aviles FX, Chang JY, and Ventura S. Folding of small disulfide-rich proteins: clarifying the puzzle. *Trends Biochem Sci* 31: 292–301, 2006.
4. Arolas JL, Bronsoms S, Aviles FX, Ventura S, and Sommerhoff CP. Oxidative folding of leech-derived tryptase

inhibitor *via* native disulfide-bonded intermediates. *Antioxid Redox Signal* 10: 77–85, 2008.

5. Arolas JL, Bronsoms S, Lorenzo J, Aviles FX, Chang JY, and Ventura S. Role of kinetic intermediates in the folding of leech carboxypeptidase inhibitor. *J Biol Chem* 279: 37261–37270, 2004.
6. Arolas JL, Bronsoms S, Ventura S, Aviles FX, and Calvete JJ. Characterizing the tick carboxypeptidase inhibitor: molecular basis for its two-domain nature. *J Biol Chem* 281: 22906–22916, 2006.
7. Arolas JL, Castillo V, Bronsoms S, Aviles FX, and Ventura S. Designing out disulfide bonds of leech carboxypeptidase inhibitor: implications for its folding, stability and function. *J Mol Biol* 392: 529–546, 2009.
8. Arolas JL, Lorenzo J, Rovira A, Vendrell J, Aviles FX, and Ventura S. Secondary binding site of the potato carboxypeptidase inhibitor. Contribution to its structure, folding, and biological properties. *Biochemistry* 43: 7973–7982, 2004.
9. Arolas JL and Ventura S. Protease inhibitors as models for the study of oxidative folding. *Antioxid Redox Signal* 14: 97–112, 2011.
10. Aslund F and Beckwith J. Bridge over troubled waters: sensing stress by disulfide bond formation. *Cell* 96: 751–753, 1999.
11. Belli M, Ramazzotti M, and Chiti F. Prediction of amyloid aggregation *in vivo*. *EMBO Rep* 12: 657–663, 2011.
12. Bulaj G and Goldenberg DP. Phi-values for BPTI folding intermediates and implications for transition state analysis. *Nat Struct Biol* 8: 326–330, 2001.
13. Calloni G, Zoffoli S, Stefani M, Dobson CM, and Chiti F. Investigating the effects of mutations on protein aggregation in the cell. *J Biol Chem* 280: 10607–10613, 2005.
14. Castillo V, Grana-Montes R, Sabate R, and Ventura S. Prediction of the aggregation propensity of proteins from the primary sequence: aggregation properties of proteomes. *Biotechnol J* 6: 674–685, 2011.
15. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, and Brenner SE. The ASTRAL Compendium in 2004. *Nucleic Acids Res* 32: D189–D192, 2004.
16. Chandonia JM, Walker NS, Lo Conte L, Koehl P, Levitt M, and Brenner SE. ASTRAL compendium enhancements. *Nucleic Acids Res* 30: 260–263, 2002.
17. Chang JY. The disulfide folding pathway of tick anticoagulant peptide (TAP), a Kunitz-type inhibitor structurally homologous to BPTI. *Biochemistry* 35: 11702–11709, 1996.
18. Chang JY. A two-stage mechanism for the reductive unfolding of disulfide-containing proteins. *J Biol Chem* 272: 69–75, 1997.
19. Chang JY. Evidence for the underlying cause of diversity of the disulfide folding pathway. *Biochemistry* 43: 4522–4529, 2004.
20. Chang JY. Diversity of folding pathways and folding models of disulfide proteins. *Antioxid Redox Signal* 10: 171–177, 2008.
21. Chang JY and Li L. 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, 2005.
22. Chang JY, Schindler P, and Chatrenet B. The disulfide structures of scrambled hirudins. *J Biol Chem* 270: 11992–11997, 1995.
23. Chatrenet B and Chang JY. The disulfide folding pathway of hirudin elucidated by stop/go folding experiments. *J Biol Chem* 268: 20988–20996, 1993.
24. Cheek S, Krishna SS, and Grishin NV. Structural classification of small, disulfide-rich protein domains. *J Mol Biol* 359: 215–237, 2006.
25. Chiti F and Dobson CM. Protein misfolding, functional amyloid, and human disease. *Annu Rev Biochem* 75: 333–366, 2006.
26. Chiti F, Stefani M, Taddei N, Ramponi G, and Dobson CM. Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature* 424: 805–808, 2003.
27. Conchillo-Solé O, de Groot NS, Avilés FX, Vendrell J, Daura X, and Ventura S. AGGRESCAN: a server for the prediction and evaluation of ''hot spots'' of aggregation in polypeptides. *BMC Bioinformatics* 8: 65, 2007.
28. This reference has been deleted.
29. Covaleda G, del Rivero MA, Chavez MA, Aviles FX, and Reverter D. Crystal structure of novel metallocarboxypeptidase inhibitor from marine mollusk *Nerita versicolor* in complex with human carboxypeptidase A4. *J Biol Chem* 287: 9250–9258, 2012.
30. Creighton TE. Intermediates in the refolding of reduced ribonuclease A. *J Mol Biol* 129: 411–431, 1979.
31. Creighton TE. Protein folding coupled to disulphide bond formation. *Biol Chem* 378: 731–744, 1997.
32. Creighton TE, Hillson DA, and Freedman RB. Catalysis by protein-disulphide isomerase of the unfolding and refolding of proteins with disulphide bonds. *J Mol Biol* 142: 43–62, 1980.
33. de Groot N, Pallares I, Aviles F, Vendrell J, and Ventura S. Prediction of ''hot spots'' of aggregation in disease-linked polypeptides. *BMC Struct Biol* 5: 18, 2005.
34. de Groot NS, Castillo V, Grana-Montes R, and Ventura S. AGGRESCAN: method, application, and perspectives for drug design. *Methods Mol Biol* 819: 199–220, 2012.
35. de Groot NS and Ventura S. Protein aggregation profile of the bacterial cytosol. *PLoS One* 5: e9383, 2010.
36. Denoncin K and Collet JF. Disulfide bond formation in the bacterial periplasm: major achievements and challenges ahead. *Antioxid Redox Signal* 19: 63–71, 2013.
37. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, and Obradovic Z. Intrinsically disordered protein. *J Mol Graph Model* 19: 26–59, 2001.
38. Espargaro A, Castillo V, de Groot NS, and Ventura S. The *in vivo* and *in vitro* aggregation properties of globular proteins correlate with their conformational stability: the SH3 case. *J Mol Biol* 378: 1116–1131, 2008.
39. Fersht AR. The sixth Datta Lecture. Protein folding and stability: the pathway of folding of barnase. *FEBS Lett* 325: 5–16, 1993.
40. Fischer M, Horn S, Belkacemi A, Kojer K, Petrungaro C, Habich M, Ali M, Kuttner V, Bien M, Kauff F, Dengjel J, Herrmann JM, and Riemer J. Protein import and oxidative folding in the mitochondrial intermembrane space of intact mammalian cells. *Mol Biol Cell* 24: 2160–2170, 2013.
41. Frand AR, Cuozzo JW, and Kaiser CA. Pathways for protein disulphide bond formation. *Trends Cell Biol* 10: 203–210, 2000.

41a. Fuller E, Green BR, Catlin P, Buczek O, Nielsen JS, Olivera BM, Bulaj G. Oxidative folding of conotoxins sharing an identical disulfide bridging framework. *FEBS J* 272: 1727–1738, 2005.

42. Galzitskaya OV, Garbuzynskiy SO, and Lobanov MY. FoldUnfold: web server for the prediction of disordered

regions in protein chain. *Bioinformatics* 22: 2948–2949, 2006.

43. Grana-Montes R, de Groot NS, Castillo V, Sancho J, Velazquez-Campoy A, and Ventura S. Contribution of disulfide bonds to stability, folding and amyloid fibril formation: the PI3-SH3 domain case. *Antioxid Redox Signal* 16: 1–15, 2012.
44. Guerois R, Nielsen JE, and Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol* 320: 369–387, 2002.
45. Guijarro JI, Sunde M, Jones JA, Campbell ID, and Dobson CM. Amyloid fibril formation by an SH3 domain. *Proc Natl Acad Sci U S A* 95: 4224–4228, 1998.
46. Guo ZY, Qiao ZS, and Feng YM. The *in vitro* oxidative folding of the insulin superfamily. *Antioxid Redox Signal* 10: 127–139, 2008.
47. Hegyi H and Tompa P. Intrinsically disordered proteins display no preference for chaperone binding *in vivo*. *PLoS Comput Biol* 4: e1000017, 2008.
48. Hogg PJ. Disulfide bonds as switches for protein function. *Trends Biochem Sci* 28: 210–214, 2003.
49. Ivanova MI, Thompson MJ, and Eisenberg D. A systematic screen of beta(2)-microglobulin and insulin for amyloid-like segments. *Proc Natl Acad Sci U S A* 103: 4079–4082, 2006.
50. Jamroz M, Kolinski A, and Kmiecik S. CABS-flex: server for fast simulation of protein structure fluctuations. *Nucleic Acids Res* 41: W427–W431, 2013.
51. Jansens A, van Duijn E, and Braakman I. Coordinated nonvectorial folding in a newly synthesized multidomain protein. *Science* 298: 2401–2403, 2002.
52. Kallberg Y, Gustafsson M, Persson B, Thyberg J, and Johansson J. Prediction of amyloid fibril-forming proteins. *J Biol Chem* 276: 12945–12950, 2001.
53. Li YJ, Rothwarf DM, and Scheraga HA. Mechanism of reductive protein unfolding. *Nat Struct Biol* 2: 489–494, 1995.
54. Linding R, Schymkowitz J, Rousseau F, Diella F, and Serrano L. A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins. *J Mol Biol* 342: 345–353, 2004.
55. Lindorff-Larsen K, Piana S, Dror RO, and Shaw DE. How fast-folding proteins fold. *Science* 334: 517–520, 2011.
56. Lobanov MY, Suvorina MY, Dovidchenko NV, Sokolovskiy IV, Surin AK, and Galzitskaya OV. A novel web server predicts amino acid residue protection against hydrogen-deuterium exchange. *Bioinformatics* 29: 1375–1381, 2013.
57. Luo X, Jiang Q, Song G, Liu YL, Xu ZG, and Guo ZY. Efficient oxidative folding and site-specific labeling of human hepcidin to study its interaction with receptor ferroportin. *FEBS J* 279: 3166–3175, 2012.
58. Mamathambika BS and Bardwell JC. Disulfide-linked protein folding pathways. *Annu Rev Cell Dev Biol* 24: 211–235, 2008.
59. Marino SM and Gladyshev VN. Cysteine function governs its conservation and degeneration and restricts its utilization on protein surfaces. *J Mol Biol* 404: 902–916, 2010.
60. Maurer-Stroh S, Debulpaep M, Kuemmerer N, Lopez de la Paz M, Martins IC, Reumers J, Morris KL, Copland A, Serpell L, Serrano L, Schymkowitz JW, and Rousseau F. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat Methods* 7: 237–242.
61. This reference has been deleted.
62. Monsellier E and Chiti F. Prevention of amyloid-like aggregation as a driving force of protein evolution. *EMBO Rep* 8: 737–742, 2007.
63. Mossuto MF, Bolognesi B, Guixer B, Dhulesia A, Agostini F, Kumita JR, Tartaglia GG, Dumoulin M, Dobson CM, and Salvatella X. Disulfide bonds reduce the toxicity of the amyloid fibrils formed by an extracellular protein. *Angew Chem Int Ed Engl* 50: 7048–7051, 2011.
64. Mu Y, Nordenskiold L, and Tam JP. Folding, misfolding, and amyloid protofibril formation of WW domain FBP28. *Biophys J* 90: 3983–3992, 2006.
65. Neklesa TK and Crews CM. Chemical biology: greasy tags for protein removal. *Nature* 487: 308–309, 2012.
66. Pace CN, Grimsley GR, Thomson JA, and Barnett BJ. Conformational stability and activity of ribonuclease T1 with zero, one, and two intact disulfide bonds. *J Biol Chem* 263: 11820–11825, 1988.
67. Pantoja-Uceda D, Arolas JL, Aviles FX, Santoro J, Ventura S, and Sommerhoff CP. Deciphering the structural basis that guides the oxidative folding of leech-derived tryptase inhibitor. *J Biol Chem* 284: 35612–35620, 2009.
68. This reference has been deleted.
69. Polticelli F, Raybaudi-Massilia G, and Ascenzi P. Structural determinants of mini-protein stability. *Biochem Mol Biol Educ* 29: 16–20, 2001.
70. Prilusky J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O, Beckmann JS, Silman I, and Sussman JL. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* 21: 3435–3438, 2005.
71. This reference has been deleted.
72. Rezaei-Ghaleh N, Blackledge M, and Zweckstetter M. Intrinsically disordered proteins: from sequence and conformational properties toward drug discovery. *Chembiochem* 13: 930–950, 2012.
73. Rousseau F, Schymkowitz J, and Serrano L. Protein aggregation and amyloidosis: confusion of the kinds? *Curr Opin Struct Biol* 16: 118–126, 2006.
74. Safavi-Hemami H, Gorasia DG, Steiner AM, Williamson NA, Karas JA, Gajewiak J, Olivera BM, Bulaj G, and Purcell AW. Modulation of conotoxin structure and function is achieved through a multienzyme complex in the venom glands of cone snails. *J Biol Chem* 287: 34288–34303, 2012.
75. Salinas G, Pellizza L, Margenat M, Flo M, and Fernandez C. Tuned *Escherichia coli* as a host for the expression of disulfide-rich proteins. *Biotechnol J* 6: 686–699, 2011.
76. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, and Serrano L. The FoldX web server: an online force field. *Nucleic Acids Res* 33: W382–W388, 2005.
77. Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, and Dunker AK. DisProt: the Database of Disordered Proteins. *Nucleic Acids Res* 35: D786–D793, 2007.
78. Siekierska A, De Baets G, Reumers J, Gallardo R, Rudyak S, Broersen K, Couceiro J, Van Durme J, Schymkowitz J, and Rousseau F. α-Galactosidase aggregation is a determinant of pharmacological chaperone efficacy on Fabry disease mutants. *J Biol Chem* 287: 28386–28397, 2012.
79. Tzotzos S and Doig AJ. Amyloidogenic sequences in native protein structures. *Protein Sci* 19: 327–348, 2010.
80. Uversky VN, Gillespie JR, and Fink AL. Why are ‘‘natively unfolded’’ proteins unstructured under physiologic conditions? *Proteins* 41: 415–427, 2000.

81. Van Durme J, Maurer-Stroh S, Gallardo R, Wilkinson H, Rousseau F, and Schymkowitz J. Accurate prediction of DnaK-peptide binding *via* homology modelling and experimental data. *PLoS Comput Biol* 5: e1000475, 2009.

82. Ventura S, Zurdo J, Narayanan S, Parreno M, Mangues R, Reif B, Chiti F, Giannoni E, Dobson CM, Aviles FX, and Serrano L. Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case. *Proc Natl Acad Sci U S A* 101: 7258–7263, 2004.

83. Wedemeyer WJ, Welker E, Narayan M, and Scheraga HA. Disulfide bonds and protein folding. *Biochemistry* 39: 4207–4216, 2000.

84. Weissman JS and Kim PS. Reexamination of the folding of BPTI: predominance of native intermediates. *Science* 253: 1386–1393, 1991.

85. Yang ZR, Thomson R, McNeil P, and Esnouf RM. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* 21: 3369–3376, 2005.

Address correspondence to:
Dr. Salvador Ventura
Departament de Bioquimica i Biologia Molecular
Institut de Biotecnologia i Biomedicina
Universitat Autònoma de Barcelona
Barcelona 08193
Spain

E-mail: salvador.ventura@uab.es

Date of first submission to ARS Central, July 17, 2013; date of final revised submission, March 6, 2014; date of acceptance, March 16, 2014.

Abbreviations Used

APRs =	aggregation-prone regions
ATR-FTIR =	attenuated total reflectance–Fourier transform infrared spectroscopy
BPTI =	bovine pancreatic trypsin inhibitor
CR =	Congo red
DRDs =	disulfide-rich domains
DTT =	DL-dithiothreitol
ER =	endoplasmic reticulum
HFIP =	1,1,1,3,3,3-hexafluoro-2-propanol
IDPs =	intrinsically disordered proteins
IGF-1 =	insulin-like growth factor-1
LDL =	low density lipoprotein
LDTI =	leech-derived trypsin inhibitor
NvCI =	*Nerita versicolor* carboxypeptidase inhibitor
RONN =	regional order neural network
TAP =	tick anticoagulant peptide
TCEP =	*tris*(2-carboxyethyl)phosphine
TEM =	transmission electron microscopy
TFA =	trifluoroacetic acid
Th-T =	Thioflavin-T
wt =	wild-type

CHAPTER 6.- CONCLUDING REMARKS

I) Impact of Disulfide Cross-linking on the Formation and Toxicity of Amyloid Fibrils

- Crosslinking the N and C termini of the PI3-SH3 domain results in one of the greatest stabilizations yet observed for a protein domain, and particularly for an SH3 domain.

- Crosslinking the PI3-SH3 domain by disulfide bonding increases remarkably its stability, but to a lesser extent than expected according to theoretical models for the configurational entropy of polymeric chains. Analogous results from the disulfide crosslinking of different proteins indicate that enthalpic compensation in the unfolded state is an important contributor to the energetic balance upon disulfide bonding.

- Stabilization of the PI3-SH3 domain by disulfide crosslinking results mainly from the increase of its folding rate, which is attributed entirely to the reduction of the configurational entropy in the unfolded state. Φ-analysis indicates the N and C termini are not structured in the TS. Additionally, disulfide crosslinking has only a minor effect on the unfolding constant, thus suggesting the region formed by the N and C termini is not the most relevant for the stability of the native structure.

- Crosslinking the N and C terminii of the PI3-SH3 domain does not preclude its aggregation into amyloid-like fibrils, but strongly influences the kinetics of the reaction. Disulfide bonding affects both the nucleation and elongation constants, likely because, in the context of a NCC mechanism, the disulfide bond reduces the conformational flexibility required for structural reorganization during nucleation and for docking of oligomeric subunits during elongation.

- Disulfide bonding also affects the properties of the fibrils formed by the PI3-SH3 domain, which become shorter, more stable, with an increased stability and content of β-conformation, and a lower exposure of hydrophobic residues. Most remarkably the fibrils of the crosslinked PI3-SH3 domain are less cytotoxic, presumably because of the decreased hydrophic exposure reduced the probability of spurious interactions, and the higher stability of the fibrils lowering the population of oligomeric species.

- Extracellular SH3 domains display disulfide bonds naturally, and exhibit a higher intrinsic aggregation propensity than their intracellular counterparts. This observation is consistent with recent findings showing extracellular proteins with disulfide bonds present a significantly higher aggregation propensity. Together, our results suggest that naturally engineering disulfide bonds may act as an effective strategy to overcome protein aggregation, particularly for proteins that function under harsh environments such as the extracellular space. At the same time, the stability provided by disulfide crosslinking allows these proteins to harbor an increased aggregational load.

II) IDPRs acting as Entropic Bristles against Aggregation

- Deletion of the Nter unstructured extension of the SUMO proteins does not have any noticeable effect on its stability, nor on the conformation of the globular Ubl domain. In contrast, it influences strongly its hydrodynamic properties and hydrophobic exposure.

- The Nter unstructured extension of the SUMO domains slows down significantly their kinetics of aggregation into amyloid-like fibrils upon destabilization of the globular structure. The anti-aggregational behavior of the SUMO Nter tail results from its ability to create an excluded volume that effectively shields an exposed hydrophobic patch with high amyloidogenic propensity.

- The anti-aggregational capacity of the SUMO Nter extension, together with its compositional bias, indicates the primary function of this tails is to serve as “entropic bristle” domains in SUMO proteins.

- The function of SUMO Nter extensions as “entropic bristles” is further supported by its ability to reduce the aggregation of the Aβ42 peptide in vivo, when translationally fused at its N terminus. This shows that, in contrast to the classical description of “entropic bristle” domains being tens of hundreds of residues long, relatively short amino acid stretches can effectively exert an anti-aggregational “entropic bristle” function in vivo.

- As in the SUMO case, the presence of disordered segments is observed frequently at the N and C termini of globular proteins. The strikingly low tendency to aggregate of the patterns detected within these regions suggests they encode a potential role as anti-aggregational “entropic bristles”.

III) Protein Abundance as a Modulator of Aggregative Risk in the Cellular Environment

- The assessment of the previously observed correlation between mRNA levels and predicted aggregation propensity by employing experimental protein solubility data confirms this trend is maintained in the *E. coli* proteome, with low and high abundant mRNAs exhibiting significantly different solubility distributions of their corresponding proteins.

- Nonetheless, the relationship between protein expression and aggregation propensity is better noticed when considering real protein abundance. In this case, the difference in the solubility distribution between low and highly abundant proteins ins considerably more significant.

- A detailed analysis of their differential solubility distributions reveals highly transcripted/translated proteins are inherently highly soluble, while for low transcripted/translated polypeptides a bimodal distribution exists with an aggregation-prone population and another one significantly soluble.

- mRNA levels correlate better with protein levels for higly abundant than for low abundant ones. In a similar way, the same correlation is more significant when considering highly soluble proteins than the remaining polypeptides

- Together, these results indicate protein levels are tightly regulated in the cell on the basis of their aggregative properties. Abundant proteins appear to be under an increased selective pressure for high solubility and their synthesis is optimized at the gene expression level to yield high amounts of protein. Conversely, within low abundant proteins two populations are distinguished: aggregation-prone proteins which experience a reduced pressure against aggregation, and soluble proteins. The latter may correspond to proteins which experience dramatic variations of their cellular concentrations, thus being also under selection for increased solubility. In both cases, the regulation of their cellular abundance appears to be more regulated at the post-transcriptional, co-translational or post-translational levels.

IV) Analysis of Specific Aggregation Properties of Functional Classes from Different Organisms. Insights from the Kinase Complements of Proteomes

- The aggregation properties of kinase proteins follow the same trend observed for complete proteomes, with a decrease in overall aggregation propensity as organism complexity raises.
- The catalytic kinase domains retain a significant tendency to aggregate, which is above the average aggregation propensity of globular proteins. Nonetheless, their aggregative potential is “buffered” by being enclosed in the whole kinase protein. This effect might be achieved through an enrichment in IDPRs.
- Accordingly, although the aggregative potential of kinase domains varies significantly among different kinase groups, presumably due to functional constraints specific to each group, the tendency to aggregate of whole kinase proteins from different groups remains substantially low in all cases, and no evident association can be established between the aggregation propensity of kinase domains from a particular group and their associated proteins.
- The intrinsically high aggregation propensity of kinase domains appears to arise from limitations in the natural selection against aggregation caused by functional constraints. Among them, the requirement to achieve an appropriate folding, since predicted amyloidogenic stretches map to secondary structure elements of kinase domains; or the need to maintain their enzymatic activity, considering that conserved catalytic residues are found embedded or in close proximity to APRs.

V) The Relationship between Foldability and Aggregation Propensity

- In the absence of their disulfide bonds, the native structure of DRDs is generally predicted to be highly destabilized.
- Accordingly, a majority of DRDs are predicted to be intrinsically disordered when their disulfides are disrupted.
- DRDs exhibit a reduced aggregation propensity, which is below the tendency to aggregate estimated for the ensemble of globular proteins, and is similar to that found for IPDs. Consequently, DRDs present a very low number of chaperone binding sites.
- The low aggregational load presented by DRDs appears to arise from a compositional bias similar to that found in IDPs, with depletion on aggregation-promoting aliphatic hydrophobic residues and enrichment in β-breaker amino acids.
- Combining the analysis of intrinsic disorder with the prediction the aggregation propensity and the characterization of amyloidogenic stretches detected allows to rationalize the oxidative folding of DRDs whose pathway had already been determined.

DRDs exhibiting a high predicted disorder and low aggregation propensity without strong APRs fold through a hirudin-like mechanism with a high ammount of non-native disulfide intermediates; whereas disulfide-rich proteins without significant predicted disorder but exhibiting substantial aggregation propensity and amyloidogenic APRs follow the BPTI-like mechanism by populating a discrete number of native-like intermediates.

- The analysis of the aggregation properties of DRDs from their primary sequence allows to forecast the oxidative folding pathway they will follow, as exemplified by the pathway of NvCI, which has been characterized *de novo*.

- In general perspective, these results highlight the close interplay between the determinants for folding into a native structure and for aggregation into ordered structures. Since both the forces guiding both processes overlap to some extent, the efficient folding into a native structure implies the maintenance of a certain aggregational load, thus acting as a lower limit for the selective pressure against aggregation. Further selection would compromise the attainment of the native conformation unless additional stabilizing elements, such as disulfide bonds or cofactors are incorporated.

CHAPTER 7.- REFERENCES

Abeln, S. & Frenkel, D., 2008. Disordered flanks prevent peptide aggregation. *PLoS computational biology*, 4(12), p.e1000241.

Anfinsen, C.B., 1973. Principles that govern the folding of protein chains. *Science (New York, N.Y.)*, 181(4096), pp.223–30.

Arias-Moreno, X. et al., 2008. Scrambled isomers as key intermediates in the oxidative folding of ligand binding module 5 of the low density lipoprotein receptor. *The Journal of biological chemistry*, 283(20), pp.13627–37.

Arolas, J.L. et al., 2006. Folding of small disulfide-rich proteins: clarifying the puzzle. *Trends in biochemical sciences*, 31(5), pp.292–301.

Arolas, J.L. et al., 2008. Oxidative folding of leech-derived tryptase inhibitor via native disulfide-bonded intermediates. *Antioxidants & redox signaling*, 10(1), pp.77–85.

Arolas, J.L. et al., 2004. Role of kinetic intermediates in the folding of leech carboxypeptidase inhibitor. *The Journal of biological chemistry*, 279(36), pp.37261–70.

Arolas, J.L. & Ventura, S., 2011. Protease inhibitors as models for the study of oxidative folding. *Antioxidants & redox signaling*, 14(1), pp.97–112.

Auer, S. et al., 2008. A generic mechanism of emergence of amyloid protofilaments from disordered oligomeric aggregates. *PLoS computational biology*, 4(11), p.e1000222.

De Baets, G. et al., 2011. An evolutionary trade-off between protein turnover rate and protein aggregation favors a higher aggregation propensity in fast degrading proteins. *PLoS computational biology*, 7(6), p.e1002090.

Baker, D., 2000. A surprising simplicity to protein folding. *Nature*, 405(6782), pp.39–42.

Banachewicz, W. et al., 2011. Malleability of folding intermediates in the homeodomain superfamily. *Proceedings of the National Academy of Sciences of the United States of America*, 108(14), pp.5596–601.

Barghorn, S. & Mandelkow, E., 2002. Toward a unified scheme for the aggregation of tau into Alzheimer paired helical filaments. *Biochemistry*, 41(50), pp.14885–96.

Bayer, P. et al., 1998. Structure determination of the small ubiquitin-related modifier SUMO-1. *Journal of molecular biology*, 280(2), pp.275–86.

Bayro, M.J. et al., 2010. High-resolution MAS NMR analysis of PI3-SH3 amyloid fibrils: backbone conformation and implications for protofilament assembly and structure . *Biochemistry*, 49(35), pp.7474–84.

Belli, M., Ramazzotti, M. & Chiti, F., 2011. Prediction of amyloid aggregation in vivo. *EMBO reports*, 12(7), pp.657–63.

Bemporad, F. et al., 2008. Biological function in a non-native partially folded state of a protein. *The EMBO journal*, 27(10), pp.1525–35.

Bemporad, F. & Chiti, F., 2012. Protein misfolded oligomers: experimental approaches, mechanism of formation, and structure-toxicity relationships. *Chemistry & biology*, 19(3), pp.315–27.

Benetti, P.H. et al., 1998. Expression and characterization of the recombinant catalytic subunit of casein kinase II from the yeast Yarrowia lipolytica in Escherichia coli. *Protein expression and purification*, 13(3), pp.283–90.

Van den Berg, B., Ellis, R.J. & Dobson, C.M., 1999. Effects of macromolecular crowding on protein folding and aggregation. *The EMBO journal*, 18(24), pp.6927–33.

Bernstein, J. a et al., 2002. Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, 99(15), pp.9697–702.

Betz, S.F., 1993. Disulfide bonds and the stability of globular proteins. *Protein science□: a publication of the Protein Society*, 2(10), pp.1551–8.

Betz, S.F. & Pielak, G.J., 1992. Introduction of a disulfide bond into cytochrome c stabilizes a compact denatured state. *Biochemistry*, 31(49), pp.12337–44.

Bolognesi, B. et al., 2010. ANS binding reveals common features of cytotoxic amyloid species. *ACS chemical biology*, 5(8), pp.735–40.

Brandt, F. et al., 2009. The native 3D organization of bacterial polysomes. *Cell*, 136(2), pp.261–71.

Broome, B.M. & Hecht, M.H., 2000. Nature disfavors sequences of alternating polar and non-polar amino acids: implications for amyloidogenesis. *Journal of molecular biology*, 296(4), pp.961–8.

Brown, H.G. & Hoh, J.H., 1997. Entropic exclusion by neurofilament sidearms: a mechanism for maintaining interfilament spacing. *Biochemistry*, 36(49), pp.15035–40.

Bryan, A.W. et al., 2009. BETASCAN: probable beta-amyloids identified by pairwise probabilistic analysis. *PLoS computational biology*, 5(3), p.e1000333.

Bryngelson, J.D. & Wolynes, P.G., 1987. Spin glasses and the statistical mechanics of protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, 84(21), pp.7524–8.

Bucciantini, M. et al., 2002. Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature*, 416(6880), pp.507–11.

Buchner, J., 2010. Bacterial Hsp90 - desperately seeking clients. *Molecular Microbiology*, 76(3), pp.540–544.

Buck, P.M., Kumar, S. & Singh, S.K., 2013. On the role of aggregation prone regions in protein evolution, stability, and enzymatic catalysis: insights from diverse analyses. *PLoS computational biology*, 9(10), p.e1003291.

Bui, J.M., Cavalli, A. & Gsponer, J., 2008. Identification of aggregation-prone elements by using interaction-energy matrices. *Angewandte Chemie (International ed. in English)*, 47(38), pp.7267–9.

Campioni, S. et al., 2010. A causative link between the structure of aberrant protein oligomers and their toxicity. *Nature chemical biology*, 6, pp.140–147.

Carrió, M.M., Corchero, J.L. & Villaverde, A., 1998. Dynamics of in vivo protein aggregation: building inclusion bodies in recombinant bacteria. *FEMS microbiology letters*, 169(1), pp.9–15.

Carulla, N. et al., 2005. Molecular recycling within amyloid fibrils. *Nature*, 436(7050), pp.554–8.

Castillo, V. et al., 2010. Deciphering the role of the thermodynamic and kinetic stabilities of SH3 domains on their aggregation inside bacteria. *Proteomics*, 10(23), pp.4172–85.

Castillo, V. & Ventura, S., 2009. Amyloidogenic regions and interaction surfaces overlap in globular proteins related to conformational diseases. *PLoS computational biology*, 5(8), p.e1000476.

Chan, W. et al., 1996. Mutational effects on inclusion body formation in the periplasmic expression of the immunoglobulin VL domain REI. *Folding and Design*, 1(2), pp.77–89.

Chang, J., 2011. Diverse pathways of oxidative folding of disulfide proteins: underlying causes and folding models. *Biochemistry*, 50(17), pp.3414–31.

Chang, J.Y., 1997. A two-stage mechanism for the reductive unfolding of disulfide-containing proteins. *The Journal of biological chemistry*, 272(1), pp.69–75.

Chang, J.Y., 1996. The disulfide folding pathway of tick anticoagulant peptide (TAP), a Kunitz-type inhibitor structurally homologous to BPTI. *Biochemistry*, 35(36), pp.11702–9.

Chang, J.-Y., 2008. Diversity of folding pathways and folding models of disulfide proteins. *Antioxidants & redox signaling*, 10(1), pp.171–7.

Chang, J.Y., Li, L. & Lai, P.H., 2001. A major kinetic trap for the oxidative folding of human epidermal growth factor. *The Journal of biological chemistry*, 276(7), pp.4845–52.

Chang, J.Y., Schindler, P. & Chatrenet, B., 1995. The disulfide structures of scrambled hirudins. *The Journal of biological chemistry*, 270(20), pp.11992–7.

Chatrenet, B. & Chang, J.Y., 1993. The disulfide folding pathway of hirudin elucidated by stop/go folding experiments. *The Journal of biological chemistry*, 268(28), pp.20988–96.

Chatrenet, B. & Chang, J.Y., 1992. The folding of hirudin adopts a mechanism of trial and error. *The Journal of biological chemistry*, 267(5), pp.3038–43.

Cheek, S., Krishna, S.S. & Grishin, N. V, 2006. Structural classification of small, disulfide-rich protein domains. *Journal of molecular biology*, 359(1), pp.215–37.

Chen, Y. & Dokholyan, N. V, 2008. Natural selection against protein aggregation on self-interacting and essential proteins in yeast, fly, and worm. *Molecular biology and evolution*, 25(8), pp.1530–3.

Cheon, M. et al., 2007. Structural reorganisation and potential toxicity of oligomeric species formed during the assembly of amyloid fibrils. *PLoS computational biology*, 3(9), pp.1727–38.

Chiti, F., Taddei, N., et al., 2002. Kinetic partitioning of protein folding and aggregation. *Nature structural biology*, 9(2), pp.137–43.

Chiti, F. et al., 2000. Mutational analysis of the propensity for amyloid formation by a globular protein. *The EMBO journal*, 19(7), pp.1441–9.

Chiti, F. et al., 2003. Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature*, 424(6950), pp.805–8.

Chiti, F., Calamai, M., et al., 2002. Studies of the aggregation of mutant proteins in vitro provide insights into the genetics of amyloid diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 99 Suppl 4, pp.16419–26.

Chiti, F. & Dobson, C.M., 2009. Amyloid formation by globular proteins under native conditions. *Nature chemical biology*, 5(1), pp.15–22.

Chiti, F. & Dobson, C.M., 2006. Protein misfolding, functional amyloid, and human disease. *Annual review of biochemistry*, 75, pp.333–66.

Cissé, M. et al., 2011. Reversing EphB2 depletion rescues cognitive functions in Alzheimer model. *Nature*, 469(7328), pp.47–52.

Cohen, R.J. et al., 2000. Luminal contents of benign and malignant prostatic glands: correspondence to altered secretory mechanisms. *Human pathology*, 31, pp.94–100.

Cohen, S.I. a et al., 2012. From macroscopic measurements to microscopic mechanisms of protein aggregation. *Journal of molecular biology*, 421(2-3), pp.160–71.

Cohen, S.I. a et al., 2013. Proliferation of amyloid-β42 aggregates occurs through a secondary nucleation mechanism. *Proceedings of the National Academy of Sciences of the United States of America*, 110(24), pp.9758–63.

Cohen, S.N. et al., 1973. Construction of biologically functional bacterial plasmids in vitro. *Proceedings of the National Academy of Sciences of the United States of America*, 70(11), pp.3240–4.

Conchillo-Solé, O. et al., 2007. AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC bioinformatics*, 8, p.65.

Cortese, M.S., Uversky, V.N. & Dunker, a K., 2008. *Intrinsic disorder in scaffold proteins: getting more from less.*,

Creighton, T.E., 1977. Conformational restrictions on the pathway of folding and unfolding of the pancreatic trypsin inhibitor. *Journal of molecular biology*, 113(2), pp.275–93.

Creighton, T.E., 1978. Experimental studies of protein folding and unfolding. *Progress in biophysics and molecular biology*, 33(3), pp.231–97.

Creighton, T.E., 1992. *Proteins: Structures and Molecular Properties*,

Cremades, N. et al., 2012. Direct observation of the interconversion of normal and toxic forms of α-synuclein. *Cell*, 149(5), pp.1048–59.

Crespo, M.D., Simpson, E.R. & Searle, M.S., 2006. Population of on-pathway intermediates in the folding of ubiquitin. *Journal of molecular biology*, 360(5), pp.1053–66.

Dalessio, P.M. et al., 2005. Swapping core residues in homologous proteins swaps folding mechanism. *Biochemistry*, 44(8), pp.3082–90.

Dalgarno, D.C., Botfield, M.C. & Rickles, R.J., 1997. SH3 domains and drug design: ligands, structure, and biological function. *Biopolymers*, 43(5), pp.383–400.

Damaschun, G. et al., 2000. Conversion of yeast phosphoglycerate kinase into amyloid-like structure. *Proteins*, 39(3), pp.204–11.

Debès, C. et al., 2013. Evolutionary optimization of protein folding. *PLoS computational biology*, 9(1), p.e1002861.

Denoncin, K. & Collet, J.-F., 2013. Disulfide bond formation in the bacterial periplasm: major achievements and challenges ahead. *Antioxidants & redox signaling*, 19(1), pp.63–71.

Dill, K.A., 1990. Dominant forces in protein folding. *Biochemistry*, 29(31), pp.7133–55.

Dill, K.A. et al., 1995. Principles of protein folding--a perspective from simple exact models. *Protein science□: a publication of the Protein Society*, 4(4), pp.561–602.

Dobson, C.M., 2003a. Protein folding and disease: a view from the first Horizon Symposium. *Nature reviews. Drug discovery*, 2(2), pp.154–60.

Dobson, C.M., 2003b. Protein folding and misfolding. *Nature*, 426(6968), pp.884–90.

Dobson, C.M., 2001. The structural basis of protein folding and its links with human disease. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 356(1406), pp.133–45.

Doig, A.J. & Williams, D.H., 1991. Is the hydrophobic effect stabilizing or destabilizing in proteins? The contribution of disulphide bonds to protein stability. *Journal of molecular biology*, 217(2), pp.389–98.

Doyle, L., 1988. Lardaceous disease: some early reports by British authors (1722-1879). *Journal of the Royal Society of Medicine*, 81(12), pp.729–31.

DuBay, K.F. et al., 2004. Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains. *Journal of molecular biology*, 341(5), pp.1317–26.

Dunker, a K. et al., 2005. Flexible nets. The roles of intrinsic disorder in protein interaction networks. *The FEBS journal*, 272(20), pp.5129–48.

Dyson, H.J. & Wright, P.E., 2002. Coupling of folding and binding for unstructured proteins. *Current Opinion in Structural Biology*, 12, pp.54–60.

Eichner, T. & Radford, S.E., 2011. A diversity of assembly mechanisms of a generic amyloid fold. *Molecular cell*, 43(1), pp.8–18.

Eisenberg, D. & Jucker, M., 2012. The amyloid state of proteins in human diseases. *Cell*, 148(6), pp.1188–203.

Elam, J.S. et al., 2003. Amyloid-like filaments and water-filled nanotubes formed by SOD1 mutant proteins linked to familial ALS. *Nature structural biology*, 10(6), pp.461–7.

Ellis, R.J., 2001. Macromolecular crowding: obvious but underappreciated. *Trends in biochemical sciences*, 26(10), pp.597–604.

Esler, W.P. et al., 1996. Point substitution in the central hydrophobic cluster of a human beta-amyloid congener disrupts peptide folding and abolishes plaque competence. *Biochemistry*, 35(44), pp.13914–21.

Espargaró, A. et al., 2008. The in vivo and in vitro aggregation properties of globular proteins correlate with their conformational stability: the SH3 case. *Journal of molecular biology*, 378(5), pp.1116–31.

Fändrich, M. et al., 2003. Myoglobin forms amyloid fibrils by association of unfolded polypeptide segments. *Proceedings of the National Academy of Sciences of the United States of America*, 100(26), pp.15463–8.

Fändrich, M., Fletcher, M.A. & Dobson, C.M., 2001. Amyloid fibrils from muscle myoglobin. *Nature*, 410(6825), pp.165–6.

Fernandez-Escamilla, A.-M. et al., 2004. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature biotechnology*, 22(10), pp.1302–6.

Ferron, F. et al., 2006. A practical overview of protein disorder prediction methods. *Proteins*, 65(1), pp.1–14.

Fischer, M. et al., 2013. Protein import and oxidative folding in the mitochondrial intermembrane space of intact mammalian cells. *Molecular biology of the cell*, 24(14), pp.2160–70.

Flotho, A. & Melchior, F., 2013. Sumoylation: a regulatory protein modification in health and disease. *Annual review of biochemistry*, 82, pp.357–85.

Fraga, H. & Ventura, S., 2013. Oxidative folding in the mitochondrial intermembrane space in human health and disease. *International journal of molecular sciences*, 14(2), pp.2916–27.

Frand, A.R., Cuozzo, J.W. & Kaiser, C. a, 2000. Pathways for protein disulphide bond formation. *Trends in cell biology*, 10(5), pp.203–10.

Friel, C.T., Beddard, G.S. & Radford, S.E., 2004. Switching two-state to three-state kinetics in the helical protein Im9 via the optimisation of stabilising non-native interactions by design. *Journal of molecular biology*, 342(1), pp.261–73.

Frousios, K.K. et al., 2009. Amyloidogenic determinants are usually not buried. *BMC structural biology*, 9, p.44.

Fuller, E. et al., 2005. Oxidative folding of conotoxins sharing an identical disulfide bridging framework. *The FEBS journal*, 272(7), pp.1727–38.

Furukawa, Y. & O'Halloran, T. V, 2005. Amyotrophic lateral sclerosis mutations have the greatest destabilizing effect on the apo- and reduced form of SOD1, leading to unfolding and oxidative aggregation. *The Journal of biological chemistry*, 280(17), pp.17266–74.

Le Gall, T. et al., 2007. Intrinsic disorder in the Protein Data Bank. *J Biomol Struct Dyn*, 24, pp.325–342.

Garbuzynskiy, S.O., Lobanov, M.Y. & Galzitskaya, O. V, 2010. FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics (Oxford, England)*, 26(3), pp.326–32.

García-Fruitós, E. et al., 2011. Biological role of bacterial inclusion bodies: a model for amyloid aggregation. *The FEBS journal*, 278(14), pp.2419–27.

Gareau, J.R. & Lima, C.D., 2010. The SUMO pathway: emerging mechanisms that shape specificity, conjugation and recognition. *Nature reviews. Molecular cell biology*, 11(12), pp.861–71.

Gareau, J.R., Reverter, D. & Lima, C.D., 2012. Determinants of small ubiquitin-like modifier 1 (SUMO1) protein specificity, E3 ligase, and SUMO-RanGAP1 binding activities of nucleoporin RanBP2. *The Journal of biological chemistry*, 287(7), pp.4740–51.

Garel, J.R. & Baldwin, R.L., 1973. Both the fast and slow refolding reactions of ribonuclease A yield native enzyme. *Proceedings of the National Academy of Sciences of the United States of America*, 70(12), pp.3347–51.

Garel, J.R., Nall, B.T. & Baldwin, R.L., 1976. Guanidine-unfolded state of ribonuclease A contains both fast- and slow-refolding species. *Proceedings of the National Academy of Sciences of the United States of America*, 73(6), pp.1853–7.

Geiss-Friedlander, R. & Melchior, F., 2007. Concepts in sumoylation: a decade on. *Nature reviews. Molecular cell biology*, 8(12), pp.947–56.

Georgescauld, F. et al., 2011. Aggregation of the neuroblastoma-associated mutant (S120G) of the human nucleoside diphosphate kinase-A/NM23-H1 into amyloid fibrils. *Naunyn-Schmiedeberg's archives of pharmacology*, 384(4-5), pp.373–81.

Gianni, S. et al., 2003. Unifying features in protein-folding mechanisms. *Proceedings of the National Academy of Sciences of the United States of America*, 100(23), pp.13286–91.

Gidalevitz, T. et al., 2006. Progressive disruption of cellular protein folding in models of polyglutamine diseases. *Science (New York, N.Y.)*, 311(5766), pp.1471–4.

Glover, J.R. & Lindquist, S., 1998. Hsp104, Hsp70, and Hsp40: a novel chaperone system that rescues previously aggregated proteins. *Cell*, 94(1), pp.73–82.

Goldberg, A.L., 2003. Protein degradation and protection against misfolded or damaged proteins. *Nature*, 426(6968), pp.895–9.

Grantcharova, V.P. et al., 1998. Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. *Nature structural biology*, 5(8), pp.714–20.

Grantcharova, V.P., Riddle, D.S. & Baker, D., 2000. Long-range order in the src SH3 folding transition state. *Proceedings of the National Academy of Sciences of the United States of America*, 97(13), pp.7084–9.

Graña-Montes, R. et al., 2012. Contribution of disulfide bonds to stability, folding, and amyloid fibril formation: the PI3-SH3 domain case. *Antioxidants & redox signaling*, 16(1), pp.1–15.

Graña-Montes, R. et al., 2014. N-terminal protein tails act as aggregation protective entropic bristles: the SUMO case. *Biomacromolecules*, 15(4), pp.1194–203.

Gregersen, N., 2006. Protein misfolding disorders: pathogenesis and intervention. *Journal of inherited metabolic disease*, 29(2-3), pp.456–70.

Griffith, J.S., 1967. Self-replication and scrapie. *Nature*, 215(5105), pp.1043–4.

Gromiha, M.M. & Selvaraj, S., 2001. Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. *Journal of molecular biology*, 310(1), pp.27–32.

De Groot, N.S. et al., 2006. Mutagenesis of the central hydrophobic cluster in Abeta42 Alzheimer's peptide. Side-chain properties correlate with aggregation propensities. *The FEBS journal*, 273(3), pp.658–68.

De Groot, N.S., Sabate, R. & Ventura, S., 2009. Amyloids in bacterial inclusion bodies. *Trends in biochemical sciences*, 34(8), pp.408–16.

Guijarro, J.I., Sunde, M., et al., 1998. Amyloid fibril formation by an SH3 domain. *Proceedings of the National Academy of Sciences of the United States of America*, 95(8), pp.4224–8.

Guijarro, J.I., Morton, C.J., et al., 1998. Folding kinetics of the SH3 domain of PI3 kinase by real-time NMR combined with optical spectroscopy. *Journal of molecular biology*, 276(3), pp.657–67.

Guo, Z.-Y., Qiao, Z.-S. & Feng, Y.-M., 2008. The in vitro oxidative folding of the insulin superfamily. *Antioxidants & redox signaling*, 10(1), pp.127–39.

Gupta, A., Van Vlijmen, H.W.T. & Singh, J., 2004. A classification of disulfide patterns and its relationship to protein structure and function. *Protein science□: a publication of the Protein Society*, 13(8), pp.2045–58.

Hagihara, Y., Mine, S. & Uegaki, K., 2007. Stabilization of an immunoglobulin fold domain by an engineered disulfide bond at the buried hydrophobic region. *The Journal of biological chemistry*, 282(50), pp.36489–95.

Hanks, S.K. & Hunter, T., 1995. Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB journal□: official publication of the Federation of American Societies for Experimental Biology*, 9(8), pp.576–96.

Harper, J.D. & Lansbury, P.T., 1997. Models of amyloid seeding in Alzheimer's disease and scrapie: mechanistic truths and physiological consequences of the time-dependent solubility of amyloid proteins. *Annual review of biochemistry*, 66, pp.385–407.

Hartl, F.U., Bracher, A. & Hayer-Hartl, M., 2011. Molecular chaperones in protein folding and proteostasis. *Nature*, 475(7356), pp.324–32.

Haslberger, T. et al., 2007. M domains couple the ClpB threading motor with the DnaK chaperone activity. *Molecular cell*, 25(2), pp.247–60.

Hawe, A., Sutter, M. & Jiskoot, W., 2008. Extrinsic fluorescent dyes as tools for protein characterization. *Pharmaceutical research*, 25(7), pp.1487–99.

Hay, R.T., 2005. SUMO: a history of modification. *Molecular cell*, 18(1), pp.1–12.

He, B. et al., 2009. Predicting intrinsic disorder in proteins: an overview. *Cell research*, 19(8), pp.929–49.

Hecker, C.-M. et al., 2006. Specification of SUMO1- and SUMO2-interacting motifs. *The Journal of biological chemistry*, 281(23), pp.16117–27.

Hilbich, C. et al., 1992. Substitutions of hydrophobic amino acids reduce the amyloidogenicity of Alzheimer's disease beta A4 peptides. *Journal of molecular biology*, 228, pp.460–473.

Hogg, P.J., 2003. Disulfide bonds as switches for protein function. *Trends in biochemical sciences*, 28(4), pp.210–4.

Horwich, A.L. et al., 2007. Two families of chaperonin: physiology and mechanism. *Annual review of cell and developmental biology*, 23, pp.115–45.

Hubbard, T.J.P. et al., 1997. SCOP: a Structural Classification of Proteins database. *Nucleic Acids Research*, 25(1), pp.236–239.

Invernizzi, G. et al., 2012. Protein aggregation: mechanisms and functional consequences. *The international journal of biochemistry & cell biology*, 44(9), pp.1541–54.

Ishihama, Y. et al., 2008. Protein abundance profiling of the Escherichia coli cytosol. *BMC genomics*, 9, p.102.

Itzhaki, L.S., Otzen, D.E. & Fersht, a R., 1995. The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *Journal of molecular biology*, 254(2), pp.260–88.

Ivankov, D.N. et al., 2003. Contact order revisited: influence of protein size on the folding rate. *Protein science□: a publication of the Protein Society*, 12(9), pp.2057–62.

Ivanova, M.I. et al., 2004. An amyloid-forming segment of beta2-microglobulin suggests a molecular model for the fibril. *Proceedings of the National Academy of Sciences of the United States of America*, 101(29), pp.10584–9.

Jackson, S.E., 2006. Ubiquitin: a small protein folding paradigm. *Organic & biomolecular chemistry*, 4(10), pp.1845–53.

Jackson, S.E. & Fersht, A.R., 1991. Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition. *Biochemistry*, 30(43), pp.10428–35.

Jahn, T.R. & Radford, S.E., 2008. Folding versus aggregation: polypeptide conformations on competing pathways. *Archives of biochemistry and biophysics*, 469(1), pp.100–17.

Janowski, R., Kozak, M. & Jankowska, E., 2001. Human cystatin C , an amyloidogenic protein , dimerizes through three- dimensional domain. , 8(4).

Jansens, A., van Duijn, E. & Braakman, I., 2002. Coordinated nonvectorial folding in a newly synthesized multidomain protein. *Science (New York, N.Y.)*, 298(5602), pp.2401–3.

Jarrett, J.T. & Lansbury, P.T., 1993. Seeding “one-dimensional crystallization” of amyloid: A pathogenic mechanism in Alzheimer’s disease and scrapie? *Cell*, 73, pp.1055–1058.

Jiménez, J.L. et al., 2002. The protofilament structure of insulin amyloid fibrils. *Proceedings of the National Academy of Sciences of the United States of America*, 99(14), pp.9196–201.

Jin, F. & Liu, Z., 2013. Inherent relationships among different biophysical prediction methods for intrinsically disordered proteins. *Biophysical journal*, 104(2), pp.488–95.

Johnson, C.M. et al., 1997. Thermodynamics of denaturation of mutants of barnase with disulfide crosslinks. *Journal of molecular biology*, 268(1), pp.198–208.

Johnson, E.S., 2004. Protein modification by SUMO. *Annual review of biochemistry*, 73, pp.355–82.

Johnston, J. a, Ward, C.L. & Kopito, R.R., 1998. Aggresomes: a cellular response to misfolded proteins. *The Journal of cell biology*, 143(7), pp.1883–98.

Kaganovich, D., Kopito, R. & Frydman, J., 2008. Misfolded proteins partition between two distinct quality control compartments. *Nature*, 454(7208), pp.1088–95.

Kajava, A. V, Aebi, U. & Steven, A.C., 2005. The parallel superpleated beta-structure as a model for amyloid fibrils of human amylin. *Journal of molecular biology*, 348(2), pp.247–52.

Kampinga, H.H. & Craig, E. a, 2010. The HSP70 chaperone machinery: J proteins as drivers of functional specificity. *Nature reviews. Molecular cell biology*, 11(8), pp.579–92.

Karplus, M., 2011. Behind the folding funnel diagram. *Nature Chemical Biology*, 7(7), pp.401–404.

Karplus, M. & Weaver, D.L., 1994. Protein folding dynamics: the diffusion-collision model and experimental data. *Protein science□: a publication of the Protein Society*, 3(4), pp.650–68.

Karplus, M. & Weaver, D.L., 1976. Protein-folding dynamics. *Nature*, 260(5550), pp.404–6.

Kauzmann, W., 1959. Some factors in the interpretation of protein denaturation. *Advances in protein chemistry*, 14, pp.1–63.

Kayatekin, C., Zitzewitz, J. a & Matthews, C.R., 2010. Disulfide-reduced ALS variants of Cu, Zn superoxide dismutase exhibit increased populations of unfolded species. *Journal of molecular biology*, 398(2), pp.320–31.

Kayed, R. et al., 2003. Common structure of soluble amyloid oligomers implies common mechanism of pathogenesis. *Science (New York, N.Y.)*, 300(5618), pp.486–9.

Khorasanizadeh, S. et al., 1993. Folding and stability of a tryptophan-containing mutant of ubiquitin. *Biochemistry*, 32(27), pp.7054–7063.

Khorasanizadeh, S., Peters, I.D. & Roder, H., 1996. Evidence for a three-state model of protein folding from kinetic analysis of ubiquitin variants with altered core residues. *Nature Structural Biology*, 3(2), pp.193–205.

Kim, D.E., Fisher, C. & Baker, D., 2000. A breakdown of symmetry in the folding transition state of protein L. *Journal of molecular biology*, 298(5), pp.971–84.

Kim, P.S. & Baldwin, R.L., 1990. Intermediates in the folding reactions of small proteins. *Annual review of biochemistry*, 59, pp.631–60.

Kim, P.S. & Baldwin, R.L., 1982. Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. *Annual review of biochemistry*, 51, pp.459–89.

Kim, Y.E. et al., 2013. Molecular chaperone functions in protein folding and proteostasis. *Annual review of biochemistry*, 82, pp.323–55.

Kirkin, V. et al., 2009. A role for ubiquitin in selective autophagy. *Molecular cell*, 34(3), pp.259–69.

Knowles, T.P.J. et al., 2009. An analytical solution to the kinetics of breakable filament assembly. *Science (New York, N.Y.)*, 326(5959), pp.1533–7.

Knowles, T.P.J., Vendruscolo, M. & Dobson, C.M., 2014. The amyloid state and its association with protein misfolding diseases. *Nature reviews. Molecular cell biology*, 15(6), pp.384–96.

Kodali, R. & Wetzel, R., 2007. Polymorphism in the intermediates and products of amyloid assembly. *Current opinion in structural biology*, 17(1), pp.48–57.

Kolli, N. et al., 2010. Distribution and paralogue specificity of mammalian deSUMOylating enzymes. *The Biochemical journal*, 430, pp.335–344.

Kotamarthi, H.C., Sharma, R. & Koti Ainavarapu, S.R., 2013. Single-molecule studies on PolySUMO proteins reveal their mechanical flexibility. *Biophysical journal*, 104(10), pp.2273–81.

Krantz, B. a & Sosnick, T.R., 2000. Distinguishing between two-state and three-state models for ubiquitin folding. *Biochemistry*, 39(38), pp.11696–701.

Krebs, M.R.H. et al., 2004. Observation of sequence specificity in the seeding of protein amyloid fibrils. *Protein science□: a publication of the Protein Society*, 13(7), pp.1933–8.

Kremer, J.J. et al., 2000. Correlation of beta-amyloid aggregate size and hydrophobicity with decreased bilayer fluidity of model membranes. *Biochemistry*, 39(33), pp.10309–18.

Krishnan, R. & Lindquist, S.L., 2005. Structural insights into a yeast prion illuminate nucleation and strain diversity. *Nature*, 435(7043), pp.765–72.

Kuhlman, B. & Baker, D., 2000. Native protein sequences are close to optimal for their structures. *Proceedings of the National Academy of Sciences of the United States of America*, 97(19), pp.10383–8.

Kumar, A. et al., 2006. Residue-level NMR view of the urea-driven equilibrium folding transition of SUMO-1 (1-97): native preferences do not increase monotonously. *Journal of molecular biology*, 361(1), pp.180–94.

Kyle, R.A., 2001. Amyloidosis: A convoluted story. *British Journal of Haematology*, 114, pp.529–538.

Lappalainen, I., Hurley, M.G. & Clarke, J., 2008. Plasticity within the obligatory folding nucleus of an immunoglobulin-like domain. *Journal of molecular biology*, 375(2), pp.547–59.

Larson, S.M. et al., 2002. Residues participating in the protein folding nucleus do not exhibit preferential evolutionary conservation. *Journal of molecular biology*, 316(2), pp.225–33.

Larson, S.M. & Davidson, a R., 2000. The identification of conserved interactions within the SH3 domain by alignment of sequences and structures. *Protein science□: a publication of the Protein Society*, 9(11), pp.2170–80.

Last, N.B., Rhoades, E. & Miranker, A.D., 2011. Islet amyloid polypeptide demonstrates a persistent capacity to disrupt membrane integrity. *Proceedings of the National Academy of Sciences of the United States of America*, 108(23), pp.9460–5.

Lawrence, C. et al., 2010. Investigation of an anomalously accelerating substitution in the folding of a prototypical two-state protein. *Journal of molecular biology*, 403(3), pp.446–58.

Leonard, C.J., Aravind, L. & Koonin, E. V, 1998. Novel families of putative protein kinases in bacteria and archaea: evolution of the "eukaryotic" protein kinase superfamily. *Genome research*, 8(10), pp.1038–47.

Leopold, P.E., Montal, M. & Onuchic, J.N., 1992. Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proceedings of the National Academy of Sciences of the United States of America*, 89(18), pp.8721–5.

Levinthal, C., 1968. Are there pathways for protein folding? *Journal de Chimie Physique et de Physico-Chimie Biologique*, 65, pp.44–45.

Levinthal, C., 1969. How to fold graciously. *Mossbauer spectroscopy in …*, 24(41), pp.22–24.

Lewandowska, A., Matuszewska, M. & Liberek, K., 2007. Conformational properties of aggregated polypeptides determine ClpB-dependence in the disaggregation process. *Journal of molecular biology*, 371(3), pp.800–11.

Li, J., Soroka, J. & Buchner, J., 2012. The Hsp90 chaperone machinery: conformational dynamics and regulation by co-chaperones. *Biochimica et biophysica acta*, 1823(3), pp.624–35.

Li, Y.-J., Rothwarf, D.M. & Scheraga, H.A., 1995. Mechanism of reductive protein unfolding. *Nature Structural Biology*, 2(6), pp.489–494.

Liang, J. et al., 1996. Crystal structure of P13K SH3 domain at 2.0 Å resolution. *Journal of molecular biology*, pp.632–643.

Lin, S.H. et al., 1984. Influence of an extrinsic crosslink on the folding pathway of ribonuclease A. Conformational and thermodynamic analysis of crosslinked (7-lysine, 41-lysine)-ribonuclease A. *Biochemistry*, 23(23), pp.5504–5512.

Lindberg, M.O. & Oliveberg, M., 2007. Malleability of protein folding pathways: a simple reason for complex behaviour. *Current opinion in structural biology*, 17(1), pp.21–9.

Linding, R. et al., 2004. A Comparative Study of the Relationship Between Protein Structure and β-Aggregation in Globular and Intrinsically Disordered Proteins. *Journal of Molecular Biology*, 342(1), pp.345–353.

Lindorff-Larsen, K. et al., 2005. Protein folding and the organization of the protein topology universe. *Trends in biochemical sciences*, 30(1), pp.13–9.

Lindorff-Larsen, K. et al., 2004. Transition states for protein folding have native topologies despite high structural variability. *Nature structural & molecular biology*, 11(5), pp.443–9.

Lobanov, M.Y. et al., 2010. Library of disordered patterns in 3D protein structures. *PLoS computational biology*, 6(10), p.e1000958.

De Los Rios, P. et al., 2006. Hsp70 chaperones accelerate protein translocation and the unfolding of stable protein aggregates by entropic pulling. *Proceedings of the National Academy of Sciences of the United States of America*, 103(16), pp.6166–71.

Lowe, A.R. & Itzhaki, L.S., 2007. Rational redesign of the folding pathway of a modular protein. *Proceedings of the National Academy of Sciences of the United States of America*, 104(8), pp.2679–84.

Lu, P. et al., 2007. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nature biotechnology*, 25(1), pp.117–24.

Lumry, R. & Biltonen, R., 1966. Validity of the "two-state" hypothesis for conformational transitions of proteins. *Biopolymers*, 4(8), pp.917–44.

Maeda, M. et al., 1992. Molecular abnormalities of a human glucose-6-phosphate dehydrogenase variant associated with undetectable enzyme activity and immunologically cross-reacting material. *American journal of human genetics*, 51(2), pp.386–95.

Mahajan, R. et al., 1997. A small ubiquitin-related polypeptide involved in targeting RanGAP1 to nuclear pore complex protein RanBP2. *Cell*, 88, pp.97–107.

Makin, O.S. et al., 2005. Molecular basis for amyloid fibril formation and stability. *Proceedings of the National Academy of Sciences of the United States of America*, 102(2), pp.315–20.

Makin, O.S. & Serpell, L.C., 2005. Structures for amyloid fibrils. *The FEBS journal*, 272(23), pp.5950–61.

Mamathambika, B.S. & Bardwell, J.C., 2008. Disulfide-linked protein folding pathways. *Annual review of cell and developmental biology*, 24, pp.211–35.

Manning, G. et al., 2002. Evolution of protein kinase signaling from yeast to man. *Trends in biochemical sciences*, 27(10), pp.514–20.

Mannini, B. et al., 2014. Toxicity of Protein Oligomers Is Rationalized by a Function Combining Size and Surface Hydrophobicity. *ACS chemical biology*.

Marblestone, J.G. et al., 2006. Comparison of SUMO fusion technology with traditional gene fusion systems: enhanced expression and solubility with SUMO. *Protein science□: a publication of the Protein Society*, 15(1), pp.182–9.

Marin, V. et al., 2010. Characterization of neuronal Src kinase purified from a bacterial expression system. *Protein expression and purification*, 74(2), pp.289–97.

Marino, S.M. & Gladyshev, V.N., 2010. Cysteine function governs its conservation and degeneration and restricts its utilization on protein surfaces. *Journal of molecular biology*, 404(5), pp.902–16.

Martínez, J.C. & Serrano, L., 1999. The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nature structural biology*, 6(11), pp.1010–6.

Mas, J., Aloy, P. & Martí-Renom, M., 1998. Protein similarities beyond disulphide bridge topology. *Journal of molecular …*, pp.541–548.

Masino, L. et al., 2011. Functional interactions as a survival strategy against abnormal aggregation. *FASEB journal□: official publication of the Federation of American Societies for Experimental Biology*, 25(1), pp.45–54.

Matouschek, a & Fersht, a R., 1993. Application of physical organic chemistry to engineered mutants of proteins: Hammond postulate behavior in the transition state of protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, 90(16), pp.7814–8.

Matouschek, A. et al., 1989. Mapping the transition state and pathway of protein folding by protein engineering. *Nature*, 340(6229), pp.122–6.

Matthews, C.R., 1987. Effect of point mutations on the folding of globular proteins. *Methods in enzymology*, 154(1985), pp.498–511.

Matunis, M.J., 1996. A novel ubiquitin-like modification modulates the partitioning of the Ran-GTPase-activating protein RanGAP1 between the cytosol and the nuclear pore complex. *The Journal of Cell Biology*, 135(6), pp.1457–1470.

Maurer-Stroh, S. et al., 2010. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nature methods*, 7(3), pp.237–42.

McCallister, E.L., Alm, E. & Baker, D., 2000. Critical role of beta-hairpin formation in protein G folding. *Nature structural biology*, 7(8), pp.669–73.

Merlini, G. & Bellotti, V., 2003. Molecular mechanisms of amyloidosis. *The New England journal of medicine*, 349(6), pp.583–96.

Minor, D.L. & Kim, P.S., 1996. Context-dependent secondary structure formation of a designed protein sequence. *Nature*, 380(6576), pp.730–4.

Mirny, L. & Shakhnovich, E., 2001. Evolutionary conservation of the folding nucleus. *Journal of molecular biology*, 308(2), pp.123–9.

Monsellier, E. et al., 2008. Aggregation propensity of the human proteome. *PLoS computational biology*, 4(10), p.e1000199.

Monsellier, E. & Chiti, F., 2007. Prevention of amyloid-like aggregation as a driving force of protein evolution. *EMBO reports*, 8(8), pp.737–42.

Morcos, F. et al., 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America*, 108(49), pp.E1293–301.

Morel, B. et al., 2010. Environmental conditions affect the kinetics of nucleation of amyloid fibrils and determine their morphology. *Biophysical journal*, 99(11), pp.3801–10.

Morris, A.M., Watzky, M. a & Finke, R.G., 2009. Protein aggregation kinetics, mechanism, and curve-fitting: a review of the literature. *Biochimica et biophysica acta*, 1794(3), pp.375–97.

Morrow, J.F. et al., 1974. Replication and transcription of eukaryotic DNA in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 71(5), pp.1743–7.

Mossuto, M.F., 2013. Disulfide bonding in neurodegenerative misfolding diseases. *International journal of cell biology*, 2013(Table 1), p.318319.

Mossuto, M.F. et al., 2011. Disulfide bonds reduce the toxicity of the amyloid fibrils formed by an extracellular protein. *Angewandte Chemie (International ed. in English)*, 50(31), pp.7048–51.

Munishkina, L. a et al., 2004. The effect of macromolecular crowding on protein aggregation and amyloid fibril formation. *Journal of molecular recognition□: JMR*, 17(5), pp.456–64.

Namanja, A.T. et al., 2012. Insights into high affinity small ubiquitin-like modifier (SUMO) recognition by SUMO-interacting motifs (SIMs) revealed by a combination of NMR and peptide array analysis. *The Journal of biological chemistry*, 287(5), pp.3231–40.

Narayan, M. et al., 2000. Oxidative Folding of Proteins. *Accounts of Chemical Research*, 33(11), pp.805–812.

Nauli, S., Kuhlman, B. & Baker, D., 2001. Computer-based redesign of a protein folding pathway. *Nature structural biology*, 8(7), pp.602–5.

Nelson, R. et al., 2005. Structure of the cross-beta spine of amyloid-like fibrils. *Nature*, 435(7043), pp.773–8.

Nelson, R. & Eisenberg, D., 2006. Structural models of amyloid-like fibrils. *Advances in protein chemistry*, 73(06), pp.235–82.

Nickson, A. a & Clarke, J., 2010. What lessons can be learned from studying the folding of homologous proteins? *Methods (San Diego, Calif.)*, 52(1), pp.38–50.

Nickson, A. a, Wensley, B.G. & Clarke, J., 2013. Take home lessons from studies of related proteins. *Current opinion in structural biology*, 23(1), pp.66–74.

Nilsson, M.R., 2004. Techniques to study amyloid fibril formation in vitro. *Methods (San Diego, Calif.)*, 34(1), pp.151–60.

Niwa, T. et al., 2009. Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 106(11), pp.4201–6.

Oliveberg, M., 1998. Alternative Explanations for “Multistate” Kinetics in Protein Folding: Transient Aggregation and Changing Transition-State Ensembles †. *Accounts of Chemical Research*, 31(11), pp.765–772.

Olofsson, A. et al., 2004. Probing solvent accessibility of transthyretin amyloid by solution NMR spectroscopy. *The Journal of biological chemistry*, 279(7), pp.5699–707.

Olzscha, H. et al., 2011. Amyloid-like aggregates sequester numerous metastable proteins with essential cellular functions. *Cell*, 144(1), pp.67–78.

Onuchic, J.N., Luthey-Schulten, Z. & Wolynes, P.G., 1997. Theory of protein folding: the energy landscape perspective. *Annual review of physical chemistry*, 48(1), pp.545–600.

Orengo, C. a, Jones, D.T. & Thornton, J.M., 1994. Protein superfamilies and domain superfolds. *Nature*, 372(6507), pp.631–4.

Orte, A. et al., 2008. Direct characterization of amyloidogenic oligomers by single-molecule fluorescence. *Proceedings of the National Academy of Sciences of the United States of America*, 105(38), pp.14424–9.

Owerbach, D. et al., 2005. A proline-90 residue unique to SUMO-4 prevents maturation and sumoylation. *Biochemical and biophysical research communications*, 337(2), pp.517–20.

Pantoja-Uceda, D. et al., 2009. Deciphering the structural basis that guides the oxidative folding of leech-derived tryptase inhibitor. *The Journal of biological chemistry*, 284(51), pp.35612–20.

Parrini, C. et al., 2005. Glycine residues appear to be evolutionarily conserved for their ability to inhibit aggregation. *Structure (London, England□: 1993)*, 13(8), pp.1143–51.

Patki, A.U., Hausrath, A.C. & Cordes, M.H.J., 2006. High polar content of long buried blocks of sequence in protein domains suggests selection against amyloidogenic non-polar sequences. *Journal of molecular biology*, 362(4), pp.800–9.

Pawar, A.P. et al., 2005. Prediction of “aggregation-prone” and “aggregation-susceptible” regions in proteins associated with neurodegenerative diseases. *Journal of molecular biology*, 350(2), pp.379–92.

Pawson, T., 1995. Protein modules and signalling networks. *Nature*, 373(6515), pp.573–80.

Pechmann, S. et al., 2009. Physicochemical principles that regulate the competition between functional and dysfunctional association of proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 106(25), pp.10159–64.

Petkova, A.T. et al., 2002. A structural model for Alzheimer’s beta -amyloid fibrils based on experimental constraints from solid state NMR. *Proceedings of the National Academy of Sciences of the United States of America*, 99(26), pp.16742–7.

Piana, S., Lindorff-Larsen, K. & Shaw, D.E., 2013. Atomic-level description of ubiquitin folding. *Proceedings of the National Academy of Sciences of the United States of America*, 110(15), pp.5915–20.

Plakoutsi, G. et al., 2004. Aggregation of the Acylphosphatase from Sulfolobus solfataricus: the folded and partially unfolded states can both be precursors for amyloid formation. *The Journal of biological chemistry*, 279(14), pp.14111–9.

Plaxco, K.W., Simons, K.T. & Baker, D., 1998. Contact order, transition state placement and the refolding rates of single domain proteins. *Journal of molecular biology*, 277(4), pp.985–94.

Poland, D.C. & Scheraga, H.A., 1965. Statistical Mechanics of Noncovalent Bonds in Polyamino Acids. VIII. Covalent Loops in Proteins. *Biopolymers*, 3, pp.379–399.

Preissler, S. & Deuerling, E., 2012. Ribosome-associated chaperones as key players in proteostasis. *Trends in biochemical sciences*, 37(7), pp.274–83.

Prusiner, S.B., 1982. Novel proteinaceous infectious particles cause scrapie. *Science (New York, N.Y.)*, 216, pp.136–144.

Ptitsyn, O.B., 1995. How the molten globule became. *Trends in biochemical sciences*, 20(9), pp.376–9.

Ptitsyn, O.B. & Rashin, A.A., 1975. A model of myoglobin self-organization. *Biophysical chemistry*, 3(1), pp.1–20.

Radivojac, P. et al., 2007. Intrinsic disorder and functional proteomics. *Biophysical journal*, 92(5), pp.1439–56.

Ramshini, H. et al., 2011. Large proteins have a great tendency to aggregate but a low propensity to form amyloid fibrils. *PloS one*, 6(1), p.e16075.

Raschke, T.M. & Levitt, M., 2005. Nonpolar solutes enhance water structure within hydration shells while reducing interactions between them. *Proceedings of the National Academy of Sciences of the United States of America*, 102(19), pp.6777–82.

Reumers, J. et al., 2009. Protein sequences encode safeguards against aggregation. *Human mutation*, 30(3), pp.431–7.

Reverter, D. & Lima, C.D., 2006. Structural basis for SENP2 protease interactions with SUMO precursors and conjugated substrates. *Nature structural & molecular biology*, 13, pp.1060–1068.

Reynolds, N.P. et al., 2011. Mechanism of membrane interaction and disruption by α-synuclein. *Journal of the American Chemical Society*, 133(48), pp.19366–75.

Richardson, J.S. & Richardson, D.C., 2002. Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proceedings of the National Academy of Sciences of the United States of America*, 99(5), pp.2754–9.

Ritter, C. et al., 2005. Correlation of structural elements and infectivity of the HET-s prion. *Nature*, 435(7043), pp.844–8.

Robinson, L.H., 1976. Psychiatric consultation for physically ill children. *Primary care*, 3(3), pp.563–75.

Rousseau, F., Serrano, L. & Schymkowitz, J.W.H., 2006. How evolutionary pressure against protein aggregation shaped chaperone specificity. *Journal of molecular biology*, 355(5), pp.1037–47.

Rubinsztein, D.C., 2006. The roles of intracellular protein-degradation pathways in neurodegeneration. *Nature*, 443(7113), pp.780–6.

Rüdiger, S. et al., 2000. Modulation of substrate specificity of the DnaK chaperone by alteration of a hydrophobic arch. *Journal of molecular biology*, 304(3), pp.245–51.

Sabate, R. et al., 2012. Native structure protects SUMO proteins from aggregation into amyloid fibrils. *Biomacromolecules*, 13(6), pp.1916–26.

Sabaté, R. et al., 2010. The role of protein sequence and amino acid composition in amyloid formation: scrambling and backward reading of IAPP amyloid fibrils. *Journal of molecular biology*, 404(2), pp.337–52.

Sabate, R., de Groot, N.S. & Ventura, S., 2010. Protein folding and aggregation in bacteria. *Cellular and molecular life sciences□: CMLS*, 67(16), pp.2695–715.

Safavi-Hemami, H. et al., 2012. Modulation of conotoxin structure and function is achieved through a multienzyme complex in the venom glands of cone snails. *The Journal of biological chemistry*, 287(41), pp.34288–303.

Saitoh, H. & Hinchey, J., 2000. Functional Heterogeneity of Small Ubiquitin-related Protein Modifiers SUMO-1 versus SUMO-2/3. *Journal of Biological Chemistry*, 275(9), pp.6252–6258.

Sakai, M. et al., 1969. Studies of corpora amylacea. I. Isolation and preliminary characterization by chemical and histochemical techniques. *Archives of neurology*, 21(5), pp.526–44.

Salamanca, S. et al., 2003. Major kinetic traps for the oxidative folding of leech carboxypeptidase inhibitor. *Biochemistry*, 42(22), pp.6754–61.

Sambashivan, S. et al., 2005. Amyloid-like fibrils of ribonuclease A with three-dimensional domain-swapped and native-like structure. *Nature*, 437(7056), pp.266–9.

Sanchez de Groot, N. et al., 2012. Evolutionary selection for protein aggregation. *Biochemical Society transactions*, 40(5), pp.1032–7.

Sánchez, I.E. et al., 2006. Point mutations in protein globular domains: contributions from function, stability and misfolding. *Journal of molecular biology*, 363(2), pp.422–32.

Sanchez-Ruiz, J.M., 2010. Protein kinetic stability. *Biophysical chemistry*, 148(1-3), pp.1–15.

Santner, A. a et al., 2012. Sweeping away protein aggregation with entropic bristles: intrinsically disordered protein fusions enhance soluble expression. *Biochemistry*, 51(37), pp.7250–62.

Scheeff, E.D. & Bourne, P.E., 2005. Structural evolution of the protein kinase-like superfamily. *PLoS computational biology*, 1(5), p.e49.

Schmidt, F.R., 2004. Recombinant expression systems in the pharmaceutical industry. *Applied microbiology and biotechnology*, 65(4), pp.363–72.

Schwartz, R., Istrail, S. & King, J., 2001. Frequencies of amino acid strings in globular protein sequences indicate suppression of blocks of consecutive hydrophobic residues. *Protein science□: a publication of the Protein Society*, 10(5), pp.1023–31.

Selkoe, D.J., 2003. Folding proteins in fatal ways. *Nature*, 426(6968), pp.900–4.

Serag, A. a et al., 2002. Arrangement of subunits and ordering of beta-strands in an amyloid sheet. *Nature structural biology*, 9(10), pp.734–9.

Serio, T.R., 2000. Nucleated Conformational Conversion and the Replication of Conformational Information by a Prion Determinant. *Science*, 289(5483), pp.1317–1321.

Serpell, L.C. et al., 2000. The protofilament substructure of amyloid fibrils. *Journal of molecular biology*, 300(5), pp.1033–9.

Shen, L. et al., 2006. SUMO protease SENP1 induces isomerization of the scissile peptide bond. *Nature structural & molecular biology*, 13, pp.1069–1077.

De Simone, A. et al., 2012. Intrinsic disorder modulates protein self-assembly and aggregation. *Proceedings of the National Academy of Sciences of the United States of America*, 109(18), pp.6951–6.

Sipe, J.D. et al., 2012. Amyloid fibril protein nomenclature: 2012 recommendations from the Nomenclature Committee of the International Society of Amyloidosis. *Amyloid□: the international journal of experimental and clinical investigation□: the official journal of the International Society of Amyloidosis*, 19(4), pp.167–70.

Sipe, J.D. & Cohen, a S., 2000. Review: history of the amyloid fibril. *Journal of structural biology*, 130(2-3), pp.88–98.

De Sousa Abreu, R. et al., 2009. Global signatures of protein and mRNA expression levels. *Molecular bioSystems*, 5(12), pp.1512–26.

Southworth, D.R. & Agard, D. a, 2011. Client-loading conformation of the Hsp90 molecular chaperone revealed in the cryo-EM structure of the human Hsp90:Hop complex. *Molecular cell*, 42(6), pp.771–81.

Spolar, R.S. & Record, M.T., 1994. Coupling of local folding to site-specific binding of proteins to DNA. *Science (New York, N.Y.)*, 263, pp.777–784.

Stefani, M., 2010. Biochemical and biophysical features of both oligomer/fibril and cell membrane in amyloid cytotoxicity. *The FEBS journal*, 277(22), pp.4602–13.

Stefani, M. & Dobson, C.M., 2003. Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution. *Journal of molecular medicine (Berlin, Germany)*, 81(11), pp.678–99.

Steward, A., Adhya, S. & Clarke, J., 2002. Sequence conservation in Ig-like domains: the role of highly conserved proline residues in the fibronectin type III superfamily. *Journal of molecular biology*, 318(4), pp.935–40.

Sun, T. et al., 2014. An antifreeze protein folds with an interior network of more than 400 semi-clathrate waters. *Science (New York, N.Y.)*, 343(6172), pp.795–8.

Sunde, M. et al., 1997. Common core structure of amyloid fibrils by synchrotron X-ray diffraction. *Journal of molecular biology*, 273(3), pp.729–39.

Sunde, M. & Blake, C., 1997. The structure of amyloid fibrils by electron microscopy and X-ray diffraction. *Advances in protein chemistry*, 50, pp.123–59.

Taipale, M. et al., 2012. Quantitative analysis of HSP90-client interactions reveals principles of substrate recognition. *Cell*, 150(5), pp.987–1001.

Taipale, M., Jarosz, D.F. & Lindquist, S., 2010. HSP90 at the hub of protein homeostasis: emerging mechanistic insights. *Nature reviews. Molecular cell biology*, 11(7), pp.515–28.

Tanaka, M. et al., 2006. The physical basis of how prion conformations determine strain phenotypes. *Nature*, 442(7102), pp.585–9.

Tanford, C., 1962. Contribution of Hydrophobic Interactions to the Stability of the Globular Conformation of Proteins. *Journal of the American Chemical Society*, 84(22), pp.4240–4247.

Tartaglia, G.G. et al., 2005. Organism complexity anti-correlates with proteomic beta-aggregation propensity. *Protein science□: a publication of the Protein Society*, 14(10), pp.2735–40.

Tartaglia, G.G. et al., 2010. Physicochemical determinants of chaperone requirements. *Journal of molecular biology*, 400(3), pp.579–88.

Tartaglia, G.G. et al., 2008. Prediction of aggregation-prone regions in structured proteins. *Journal of molecular biology*, 380(2), pp.425–36.

Tartaglia, G.G. et al., 2004. The role of aromaticity, exposed surface, and dipole moment in determining protein aggregation rates. *Protein science□: a publication of the Protein Society*, 13(7), pp.1939–41.

Tartaglia, G.G., Cavalli, A. & Vendruscolo, M., 2007. Prediction of local structural stabilities of proteins from their amino acid sequences. *Structure (London, England□: 1993)*, 15(2), pp.139–43.

Tartaglia, G.G. & Vendruscolo, M., 2009. Correlation between mRNA expression levels and protein aggregation propensities in subcellular localisations. *Molecular bioSystems*, 5(12), pp.1873–6.

Tartaglia, G.G. & Vendruscolo, M., 2008. The Zyggregator method for predicting protein aggregation propensities. *Chemical Society reviews*, 37(7), pp.1395–401.

Taylor, S.S. & Kornev, A.P., 2011. Protein kinases: evolution of dynamic regulatory proteins. *Trends in biochemical sciences*, 36(2), pp.65–77.

Tempé, D., Piechaczyk, M. & Bossis, G., 2008. SUMO under stress. *Biochemical Society transactions*, 36(Pt 5), pp.874–8.

Thompson, M.J. et al., 2006. The 3D profile method for identifying fibril-forming segments of proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 103(11), pp.4074–8.

Tiwari, A. & Hayward, L.J., 2003. Familial amyotrophic lateral sclerosis mutants of copper/zinc superoxide dismutase are susceptible to disulfide reduction. *The Journal of biological chemistry*, 278(8), pp.5984–92.

Tompa, P., 2012. Intrinsically disordered proteins: a 10-year recap. *Trends in biochemical sciences*, 37(12), pp.509–16.

Tompa, P., 2002. Intrinsically unstructured proteins. *Trends in biochemical sciences*, 27(10), pp.527–33.

Tompa, P. & Csermely, P., 2004. The role of structural disorder in the function of RNA and protein chaperones. *FASEB journal□: official publication of the Federation of American Societies for Experimental Biology*, 18(11), pp.1169–75.

Tompa, P. & Fuxreiter, M., 2008. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends in biochemical sciences*, 33(1), pp.2–8.

Trovato, A. et al., 2006. Insight into the structure of amyloid fibrils from the analysis of globular proteins. *PLoS computational biology*, 2(12), p.e170.

Tsolis, A.C. et al., 2013. A consensus method for the prediction of “aggregation-prone” peptides in globular proteins. *PloS one*, 8(1), p.e54175.

Tyedmers, J., Mogk, A. & Bukau, B., 2010. Cellular strategies for controlling protein aggregation. *Nature reviews. Molecular cell biology*, 11(11), pp.777–88.

Tzotzos, S. & Doig, A.J., 2010. Amyloidogenic sequences in native protein structures. *Protein science□: a publication of the Protein Society*, 19(2), pp.327–48.

Uversky, V.N., 2013a. A decade and a half of protein intrinsic disorder: biology still waits for physics. *Protein science□: a publication of the Protein Society*, 22(6), pp.693–724.

Uversky, V.N., 2009. Intrinsically disordered proteins and their environment: Effects of strong denaturants, temperature, pH, Counter ions, membranes, binding partners, osmolytes, and macromolecular crowding. *Protein Journal*, 28, pp.305–325.

Uversky, V.N., 2002. Natively unfolded proteins: A point where biology waits for physics. *Protein science*, pp.739–756.

Uversky, V.N., 2013b. The most important thing is the tail: multitudinous functionalities of intrinsically disordered protein termini. *FEBS letters*, 587(13), pp.1891–901.

Uversky, V.N. & Fink, A.L., 2004. Conformational constraints for amyloid fibrillation: the importance of being unfolded. *Biochimica et biophysica acta*, 1698(2), pp.131–53.

Uversky, V.N., Gillespie, J.R. & Fink, a L., 2000. Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins*, 41(3), pp.415–27.

Valdar, W.S.J., 2002. Scoring residue conservation. *Proteins*, 48(2), pp.227–41.

Ventura, S. et al., 2004. Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case. *Proceedings of the National Academy of Sciences of the United States of America*, 101(19), pp.7258–63.

Ventura, S., Lacroix, E. & Serrano, L., 2002. Insights into the Origin of the Tendency of the PI3-SH3 Domain to form Amyloid Fibrils. *Journal of Molecular Biology*, 322(5), pp.1147–1158.

Vertegaal, A.C.O. et al., 2006. Distinct and overlapping sets of SUMO-1 and SUMO-2 target proteins revealed by quantitative proteomics. *Molecular & cellular proteomics□: MCP*, 5, pp.2298–2310.

Van Vlijmen, H.W.T. et al., 2004. A Novel Database of Disulfide Patterns and its Application to the Discovery of Distantly Related Homologs. *Journal of Molecular Biology*, 335(4), pp.1083–1092.

Waldo, G. & Standish, B., 1999. Rapid protein-folding assay using green fluorescent protein. *Nature …*, pp.691–695.

Wall, J. et al., 1999. Thermodynamic Instability of Human λ6 Light Chains: Correlation with Fibrillogenicity †. *Biochemistry*, 38(42), pp.14101–14108.

Wang, L. et al., 2008. Bacterial inclusion bodies contain amyloid-like structure. *PLoS biology*, 6(8), p.e195.

Wang, L. et al., 2010. Multidimensional structure-activity relationship of a protein in its aggregated states. *Angewandte Chemie (International ed. in English)*, 49(23), pp.3904–8.

Watters, A.L. et al., 2007. The highly cooperative folding of small naturally occurring proteins is likely the result of natural selection. *Cell*, 128(3), pp.613–24.

Wedemeyer, W.J. et al., 2000. Disulfide Bonds and Protein Folding †. *Biochemistry*, 39(15), pp.4207–4216.

Weibezahn, J. et al., 2004. Thermotolerance requires refolding of aggregated proteins by substrate translocation through the central pore of ClpB. *Cell*, 119(5), pp.653–65.

Weissman, J. s. & Kim, P. s., 1991. Reexamination of the folding of BPTI: predominance of native intermediates. *Science*, 253(5026), pp.1386–1393.

Weissman, J.S. & Kim, P.S., 1992. Kinetic role of nonnative species in the folding of bovine pancreatic trypsin inhibitor. *Proceedings of the National Academy of Sciences of the United States of America*, 89(20), pp.9900–4.

Went, H.M., Benitez-Cardoza, C.G. & Jackson, S.E., 2004. Is an intermediate state populated on the folding pathway of ubiquitin? *FEBS letters*, 567(2-3), pp.333–8.

Went, H.M. & Jackson, S.E., 2005. Ubiquitin folds through a highly polarized transition state. *Protein engineering, design & selection□: PEDS*, 18(5), pp.229–37.

West, M.W. et al., 1999. De novo amyloid proteins from designed combinatorial libraries. *Proceedings of the National Academy of Sciences*, 96(20), pp.11211–11216.

Westermark, P., 2005. *Amyloid Proteins* J. D. Sipe, ed., Weinheim, Germany: Wiley-VCH Verlag GmbH.

Wetlaufer, D.B., 1973. Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins. *Proceedings of the National Academy of Sciences*, 70(3), pp.697–701.

Wilkinson, K. a & Henley, J.M., 2010. Mechanisms, regulation and consequences of protein SUMOylation. *The Biochemical journal*, 428(2), pp.133–45.

Winklhofer, K.F., Tatzelt, J. & Haass, C., 2008. The two faces of protein misfolding: gain- and loss-of-function in neurodegenerative diseases. *The EMBO journal*, 27(2), pp.336–49.

Wolde, P.R.T., 1997. Enhancement of Protein Crystal Nucleation by Critical Density Fluctuations. *Science*, 277(5334), pp.1975–1978.

Wolynes, P.G., 2008. Protein Folding, Misfolding and Aggregation. In V. Muñoz, ed. Cambridge: Royal Society of Chemistry.

Wood, S.J. et al., 1995. Prolines and amyloidogenicity in fragments of the Alzheimer's peptide beta/A4. *Biochemistry*, 34(3), pp.724–30.

Wright, C.F. et al., 2005. The importance of sequence diversity in the aggregation and evolution of proteins. *Nature*, 438(7069), pp.878–81.

Wright, P.E. & Dyson, H.J., 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of molecular biology*, 293, pp.321–331.

Wu, J., Yang, Y. & Watson, J.T., 1998. Trapping of intermediates during the refolding of recombinant human epidermal growth factor (hEGF) by cyanylation, and subsequent structural elucidation by mass spectrometry. *Protein science□: a publication of the Protein Society*, 7(4), pp.1017–28.

Wurth, C., Guimard, N.K. & Hecht, M.H., 2002. Mutations that reduce aggregation of the Alzheimer's Abeta42 peptide: an unbiased search for the sequence determinants of Abeta amyloidogenesis. *Journal of molecular biology*, 319(5), pp.1279–90.

Xie, H. et al., 2007. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *Journal of proteome research*, 6(5), pp.1882–98.

Xu, Y. et al., 2014. Structural insight into SUMO chain recognition and manipulation by the ubiquitin ligase RNF4. *Nature communications*, 5(May), p.4217.

Yam, A.Y. et al., 2008. Defining the TRiC/CCT interactome links chaperonin function to stabilization of newly made proteins with complex topologies. *Nature structural & molecular biology*, 15(12), pp.1255–62.

Yoon, S. et al., 2007. CSSP2: an improved method for predicting contact-dependent secondary structure propensity. *Computational biology and chemistry*, 31(5-6), pp.373–7.

Yoon, S. & Welsh, W.J., 2004. Detecting hidden sequence propensity for amyloid fibril formation. *Protein science□: a publication of the Protein Society*, 13(8), pp.2149–60.

Zavodszky, M. et al., 2001. Disulfide bond effects on protein stability: designed variants of Cucurbita maxima trypsin inhibitor-V. *Protein science□: a publication of the Protein Society*, 10(1), pp.149–60.

Zhu, S. et al., 2009. Protection from Isopeptidase-Mediated Deconjugation Regulates Paralog-Selective Sumoylation of RanGAP1. *Molecular Cell*, 33, pp.570–580.

Zibaee, S. et al., 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 science□: a publication of the Protein Society*, 16(5), pp.906–18.

Zietkiewicz, S., Krzewska, J. & Liberek, K., 2004. Successive and synergistic action of the Hsp70 and Hsp100 chaperones in protein disaggregation. *The Journal of biological chemistry*, 279(43), pp.44376–83.

Zimmerman, S.B. & Trach, S.O., 1991. Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of Escherichia coli. *Journal of molecular biology*, 222(3), pp.599–620.

Zuo, X. et al., 2005. Enhanced expression and purification of membrane proteins by SUMO fusion in Escherichia coli. *Journal of structural and functional genomics*, 6(2-3), pp.103–11.

CHAPTER 8.- ANNEX. STEPPING BACK AND FORWARD ON SUMO FOLDING EVOLUTION

8.1.- Introduction

Small ubiquitin-related modifiers (SUMO) participate in the post-translational modification of many polypeptides through a process known as SUMOylation, which is involved in the regulation of an increasing number of relevant cellular processes in eukaryotes including transcriptional regulation, nuclear transport, apoptosis, protein stability, maintenance of genome integrity, response to stress, signal transduction, and cell-cycle progression (Johnson 2004; Hay 2005; Geiss-Friedlander & Melchior 2007; Wilkinson & Henley 2010; Flotho & Melchior 2013). In humans, as in many other vertebrates, there exist two main paralogues - SUMO-1 and SUMO-2 -. In contrast, only one domain belonging to the SUMO family is found in either invertebrates like *Caernohabditis elegans* and *Drosophila melanogaster*, and unicellular eukaryotes such as *Saccharomyces cerevisiae* (smt3) and *Plasmodium falciparum* (pfSUMO).

We have previously employed the SUMO domains as models in order to analyze the role of terminal IDPRs in the prevention of protein aggregation since, as it has been discussed in Chapter 4, their unstructured Nter tails function as antiaggregational “entropic bristles”. We have found that SUMO domains also constitute an extremely interesting model for the study of protein folding because the analysis of the folding kinetics of human SUMO-1 and 2 reveals, quite surprisingly, that these homologous domains fold and unfold at significantly different rates while sharing the same Ubiquitin-like (Ubl) fold and owning a 43% sequence identity; thus indicating they fold through different pathways. The sequential analysis of SUMO domains from different organisms unveils a key position structurally located in the center of the extended helix, where only a strictly binary occupancy is permitted, subsequently allowing to classify the domains between SUMO-1 and SUMO-2-like sequences, the latter being already present in plants and yeasts. This suggests SUMO-2 might be more ancient and SUMO-1 to have evolved from an ancient SUMO-2-like domain. Astonishingly, exchanging the residue at that key position in SUMO-1, with that of paralogue 2, results in a shift of both the folding and unfolding kinetics to those of SUMO-2, in a movement interpreted as stepping backwards in the evolution of SUMO folding. The results presented here provide evidence of how has the folding of the SUMO domain evolved, and represent the first switch between the folding pathways of different proteins that has been achieved by means of a single substitution so far. At the same time it highlights the relevance of the comparative analysis of the folding pathways and mechanisms of homologous proteins, which has already provided extremely valuable insights to the understanding of the protein folding problem (Nickson & Clarke 2010; Nickson et al. 2013).

8.2.- RESULTS AND DISCUSSION

SUMO-1 and SUMO-2 Fold through different Pathways

The folding kinetics of the main SUMO paralogues in humans –SUMO-1 and SUMO-2– was analyzed employing the globular region of the proteins, dispensing with the Nter unstructured tails which have been previously demonstrated not to affect the stability of the domains but to act as entropic bristles, possessing a foremost anti-aggregational role (Graña-Montes et al. 2014). The wild-type (wt) domains are devoid of Trp residues and, although the Tyr signal allows to follow the folding and unfolding kinetics to a certain extent, the change of their intrinsic fluorescence quantum yield upon transition is rather small, in such a way that kinetic traces exhibit low amplitude and usually strong noise either at refolding or unfolding (Figs. 1.A and 1.B), thus requiring relatively high initial concentrations of protein (up to 100μM) so as to collect traces that could be fitted but still with substantial error. Taking into account the previous observation that SUMO domains carry a significant aggregation potential (Sabate et al. 2012; Graña-Montes et al. 2014) and considering that the use of high protein concentrations has been reported to induce artifacts when following the refolding kinetics of several proteins due to aggregation side reactions (Oliveberg 1998; Went et al. 2004), a Trp was engineered in order to improve the signal yield for further analysis of the SUMO domains folding kinetics. The Trp probe was introduced in substitution of a buried Phe, residue number 66 in SUMO-1 and 62 in SUMO-2, which correspond to the structurally equivalent position where a Trp was engineered for Ubiquitin (Ub) (Khorasanizadeh et al. 1993).

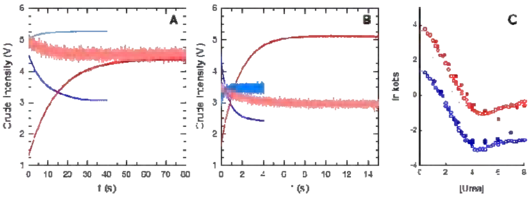


Figure 1.- Kinetic Analysis of wt SUMO Domains and Engineered Varants. Refolding (light blue) and unfolding (pink) traces of the wt domains, and refolding (blue) and unfolding (red) curves of the domains with an engineered Trp for **A)** SUMO-1 and **B)** SUMO-2. Refolding traces shown were obtained at ≈3M Urea and unfolding curves at ≈ 7M Urea. **C)** Chevron plots of wt SUMO-1 (solid blue) and SUMO-2 (solid red) and Trp bearing variants S1* (open blue) and S2* (open red).

The kinetic traces acquired for the domains engineered to endow the fluorescent probe -that will be noted S1* and S2*, corresponding to SUMO-1 F66W and SUMO-2 F62W variants, respectively- could accurately be fitted to a single exponential and its chevron plots do not show extensive curvature neither at its refolding nor at the unfolding arms (Fig 2.A), thus indicating a single reaction is taking place, which reflects these SUMO domains predominantly fold through a pathway characterized by a transition between two major states. A two-state folding pathway for the SUMO proteins is consistent with previous findings showing thermal and mechanical unfolding of human SUMO domains adjust nicely to two-state models (Sabate et al. 2012; Graña-Montes et al.

2014; Kotamarthi et al. 2013). Urea-induced denaturation of SUMO-1 has also been shown to follow a two-state scheme by intrinsic fluorescence (Kumar et al. 2006), although the authors claim SUMO-1 unfolds according to a three-state model since its Urea denaturation curve followed by far-UV CD presents two transitions.

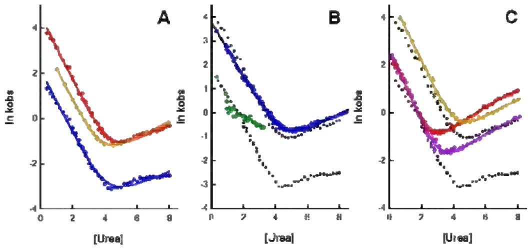


Figure 2.- Folding Kinetics of SUMO Domains. Refolding and unfolding kinetics of SUMO domains with an engineered Trp are represented in chevron plots for A) SUMO-1 (S1*) in blue, SUMO-2 (S2*) in red, and smt3 (smt3*) and their variants B) K48M (refolding constants with high -in blue- and low –in green- amplitudes) and C) M48K, in red, E73D, in ocre, and the double mutant M48K+E73D, in magenta. In B and C, S1* -lower- and S1* -upper- chevron plots are shown in black as references.

Refolding and unfolding rates are better derived from kinetic traces of the domains with an engineered Trp. However, folding rates could also be calculated, in spite of the poor signal, from kinetic traces of the wt domains, by employing high initial protein concentration, at certain Urea concentrations. Refolding and unfolding rates calculated for the wt domains present a good overlapping with the chevron plots of the engineered proteins (Fig. 1.C), so this is taken as a faithful indication that the Trp probe introduced in the proteins does not alter the folding kinetics of the SUMO domains.

Although it is already known that for homologous proteins sharing the same fold it is not a direct implication that they would fold through the same pathway or according to the same mechanism (Nickson & Clarke 2010; Nickson et al. 2013), it is still surprising to find that such a pair of closely related proteins from the same organism possess folding rates over an order of magnitude, and unfolding constants five-fold different. From the calculated kinetic constants it can be derived that SUMO-2 possesses a higher thermodynamic stability, in consonance with the previously observed higher thermal stability for SUMO-2 (Sabate et al. 2012). Nonetheless, the kinetic data also imply that SUMO-1 presents a higher energy at its TSE, relative to both the folded and the unfolded states, than that of SUMO-2. This finding suggests the main SUMO paralogues in humans fold through different pathways.

Discrete Substitutions Switch the Folding Pathway of SUMO Domains

In order to shed light on the reasons underlying such a different folding behavior, the folding kinetics of the related SUMO domain from budding yeast (smt3) was also analyzed (Fig 2.C), smt3 shows a folding constant lying closer to that of SUMO-2, being 4 times faster than SUMO-1 and a 3-fold slower than SUMO-2 (Table 1). However, its unfolding rate is much similar to that of SUMO-2, being a 5-fold faster than the rate calculated for SUMO-1.

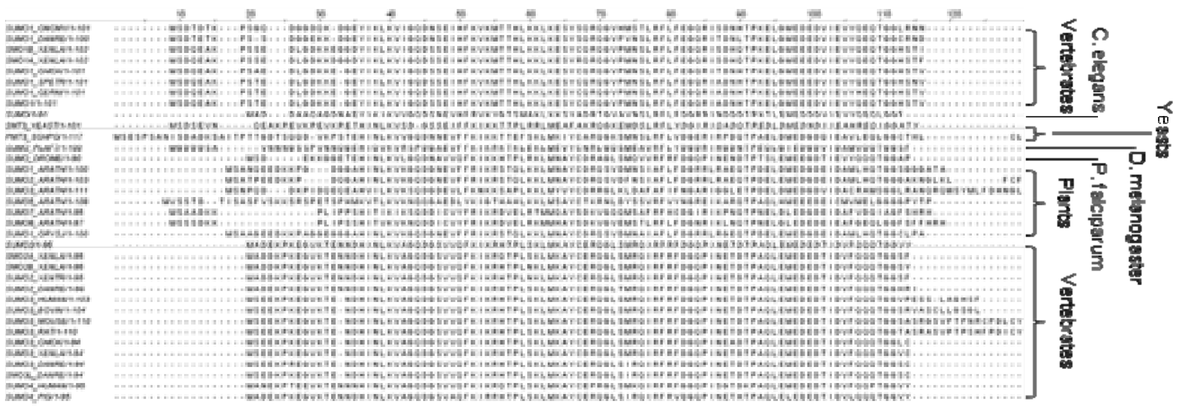


Figure 3.- Sequential Alignment of SUMO Domains from different Organisms. Alignment of non-redundant sequences retrieved from all the reviewed Uniprot annotations including "small ubiquitin-like modifier" as a keyword on its "name" field, and supplemented with SUMO sequences from *S. cerevisiae* (Q12306)*, S. pombe* (O13351)*, D melanogaster* (O97102) *and P. Falciparum* (Q8I444).

Moreover, the stabilities derived from the kinetic data indicate smt3 is as stable as SUMO-1, being SUMO-2 slightly more stable, as discussed above. In the context of the extended view suggesting that evolution shapes globular proteins to efficiently fold into stable structures, yet allowing a certain conformational flexibility to perform its functions, it may be rationalized that evolution has carved the sequence of smt3 so a faster folding is achieved, leading to slightly higher stability in SUMO-2. However, it is not straightforward to explain the action of evolution on smt3 to slow both the folding and unfolding kinetics in order to retain the same stability in SUMO-1. In the search for further explanations of the different folding pathways followed by the main SUMO paralogues in human, and how they may have evolved from a common ancestor reflected in the yeast SUMO domain, the sequences of SUMO domains from different organisms were analyzed (Fig. 3). Although the issue of the conservation of residues which are more relevant for the folding of globular proteins has remained controversial (Mirny & Shakhnovich 2001; Larson et al. 2002), likely because of the difficulty to measure the absolute contribution of specific residues to the energetics of folding (i.e. positions with non-extreme ϕ-values) as well as relative movements of the folding nucleus within the fold (Lappalainen et al. 2008), the implicit limitations of conservation scoring methods (Valdar 2002) and the possibility that no unique conservation trend might completely characterize the ensemble of the different folding mechanisms, among other reasons, the analysis of sequence evolution by means of alignment of homologous proteins may still allow to derive implications for the folding of specific domains (Lappalainen et al. 2008).

The alignment of SUMO sequences points, interestingly, to a single position which is only occupied by either a Lys or a Met in all the sequences analyzed (Fig. 3). The length of the globular, structured region of SUMO domains is almost identical in the sequences identified as such, from all the different organisms, and the matched positions in the alignment are also structurally equivalent to a large extent in the three-dimensional structures available. For clarity, therefore, a unique numbering scheme is employed throughout the text for the ensemble of SUMO domains, which corresponds to that of the human SUMO-1 sequence (UniProt accession code P63165), in order to identify structural positions and engineered mutants of the SUMO domains. According to this scheme, position 48 is where only alternaton between Lys and Met is observed, and maps to the central region of the extended helix. Considering the aminoacid occupancy at this position solely, SUMO proteins may be classified into two groups: SUMO-1-like sequences, those harboring a Lys, and SUMO-2-like sequences, presenting Met. The latter are present in yeast, plasmodium, fruit fly and different species of plants, as well as in vertebrates, which possess several SUMO-2-like domains with high sequential homology between them (SUMO-2, SUMO-3 and SUMO-4 -this one is thought to be incompetent for SUMOylation (Owerbach et al. 2005)); while SUMO-1-like sequences are only present in vertebrates and nematode. This suggest that SUMO-2-like could correspond to a more ancient form of the SUMO domain, already present in unicellular eukaryotes and plants, from which SUMO-1 like sequences evolved; whereas in vertebrates both forms have been conserved, in other pluricellular eukaryotes only one form has remained. In order to test the relevance of position 48 in the evolution of the folding of the SUMO proteins, the aminoacids at this position were exchanged for the engineered S1* and S2* domains.

Surprisingly, the K48M substitution in SUMO-1 shifts both the folding and unfolding kinetics to the rates of SUMO-2 (Fig. 2B), although the refolding reaction is best fitted to double exponentials, thus suggesting that the folding pathway of this mutant becomes more complex than that observed for the wt SUMO domains. This does not come as a surprise considering the impact the evolutionary history of a sequence has on its folding kinetics (Watters et al. 2007). Nonetheless, the derived refolding constants with the higher amplitude overlap with SUMO-2 refolding rates. Aminoacid substitutions increasing unfolding rates are the most common in the literature (Lawrence et al. 2010) but mutations increasing refolding constants by 2-fold or larger, like in the case of K48M, are rather scarce. Such a single point substitution significantly increasing both the refolding and folding rates, while maintaining the wt stability, had only been reported for a CI-2 mutant where a functionally constrained Arg was substituted by a non polar Phe (Lawrence et al. 2010). In the SUMO case, a similar principle of substituting a charged Lys by a less polar Met applies; however, in the SUMO domains the engineered mutation implies the transference of the Lys sidechain to a different aminoacid context instead of merely optimizing the polarity of the side chain to the pre-existing environment, in order to achieve a more efficient folding, as it seems to happen in the CI-2 case. The shift in the folding kinetics also implies that the energetics relative to the folded and unfolded states of the K48M main TSE is very similar to that of the SUMO-2, subsequently suggesting that this single substitution is able to shift the folding pathway of the SUMO domain. Switching the folding pathways of homologous proteins by means of discrete aminoacid substitutions is a goal that has remained largely elusive. Changes in the roughness of the pathway leading to the introduction or suppression of detectable intermediates have been achieved with a single point mutation (Dalessio et al. 2005) or by introduction a small number of substitutions (Friel et al. 2004). A shift in the flux through the parallel pathways by which an ankyrin repeat may fold was achieved by destabilizing a part of the protein with the introduction of several point mutations (Lowe & Itzhaki 2007). Perhaps the attempt which has arrived the closest to achieve the above-mentioned goal, was the change in the polarization of the TSE of protein G to that of protein L by extensive engineering (Nauli et al. 2001). To the best of our knowledge, the results presented here for the K48M mutant represent the first shift between the folding pathways of two proteins, by means of a single substitution, reported to date.

The reverse substitution where the M48K mutant was engineered in S2* slows the refolding to the SUMO-1 range, however, at the same time this mutation increases the unfolding rates relative to SUMO-2 (Fig. 2C). Although this results can be interpreted in the sense that the substitution with Lys shifts the folding pathway to that of SUMO-1 while, conversely, the introduction of a charged amino group in a largely buried position also destabilizes the native state; it may be argued as well that reduction of refolding rates arises merely due to the destabilizing effect of the Lys ε-amino in an apolar environment.

According to our hypothesis, if SUMO-2-like sequences represent a more ancient form of the SUMO domain, reverting the folding pathway of the more evolved SUMO-1 form by the introduction of a single substitution appears to have a greater chance than the reverse substitution that would imply “advancing in evolution”. In this sense, the K48M mutation allows to "step back” to the ancient pathway but at the cost of “deforming” the energy landscape. Conversely, evolving a folding pathway does not seem feasible by single aminoacid changes but would rather require the concerted action of co-ocurring substitutions which could compensate for the destabilizing effect of single mutations alone. Such compensatory substitutions that could explain the evolution of the folding pathway in the SUMO domains are not evident from a simple MSA. The aforementioned evolutive rationale has already been exploited to detect conserved pairs of interacting residues, in order to predict the three-dimensional structure of polypeptides or functional protein-protein interactions, by covariance analysis in MSAs of homologous proteins (Morcos et al. 2011). However, these approaches require a significant number of homologous, yet divergent, sequences and have, consequently, been restricted to the analyisis of proteins already present in prokaryotes – this means sufficiently diverged to allow for the disentanglement of direct from indirect correlations. Such a strategy is not feasible for the analysis of SUMO domains since there is only a limited number of available sequences, all of them belonging to eukaryotes, thus they do not present enough divergence to permit decoupling direct form indirect interactions.

Alternatively, the dynamic environment of position 48 was analyzed in the solution structures of monomeric human SUMO-1 (PDB 1A5R) and SUMO-2 (PDB 2AWT) domains in the search for contacts consistent with the notion of concertedly evolving interactions. For SUMO-1, Lys48 appears to maintain the hydrocarbon moiety of its sidechain substantially buried, while its ε-amino is exposed and close to the γ-carbonyl group of Asp73, thus these residues could establish persistent electrostatic interactions. Interestingly, Met48 is completely buried in SUMO-2 conformers and position 73 is occupied by a Glu residue, which sidechain possesses and additional aliphatic carbon compared to Asp; therefore favoring the burial of the Met apolar sidechain. In order to obtain additional insights of the interaction between positions 48 and 73 in the SUMO proteins the native state dynamics of the globular, structured region of the human domains was explored employing Molecular Dynamics (MD) simulations. The simulations show how the sidechain of Met48 is almost completely buried all along the simulation in SUMO-2, while Lys48 in SUMO-1 is slightly more exposed although its burial is quite significant. MD also reinforces the observation that the ε-amino of Lys48 is in close contact with the carboxylic group of Asp73, where the Lys N_{ζ} atom is between 3 to 4Å from the center of mass of the Asp carboxylic oxygens for the most part of the simulation (Fig. 4) and is always below 5Å from any of these oxygens, which is indicative of a persistent electrostatic interaction. In contrast the distance between the Met S atom and the carboxylic oxygens of E73 in SUMO-2 lies above 4Å in most frames of the simulation and exhibits large fluctuations.

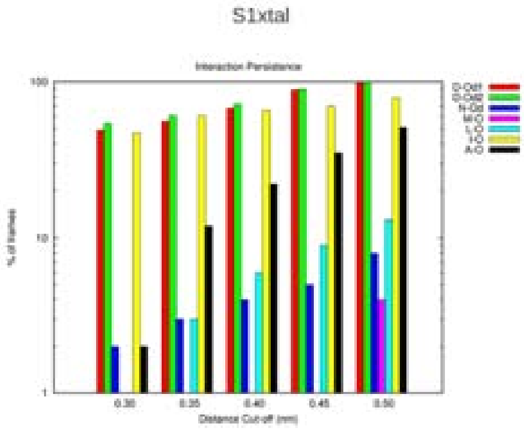


Figure 4.- Persistance (percentage of frames populated along the simulation) of the interaction between Lys48 ζ-N and different atoms in the sorrounding of position 48 for different distance cutoffs: δ-O1 and δ-O2 of Asp73, δ-O of Gln60, and backbone carbonyls of Met59, Leu62, Ile71 and Ala72.

In order to test whether the emergence of such an electrostatic interaction might in fact have driven the evolution of the SUMO domain folding, a substitution was engineered in position 73 replacing the native Glu in the SUMO-2 variant for Asp

(E73D). The addition of this substitution to M48K allows to decrease the unfolding rate by ≈3-fold while maintaining the refolding constant, whereas the E73D mutation by itself increases SUMO-2 unfolding constant (Fig. 2C). This suggests the M48K substitution actually allows to shift the SUMO folding pathway, since the concerted action of the E73D mutation allows to stabilize the folded state but has no effect on the refolding constant. However, this compensatory mutation is not enough to endow the engineered domain with a native-like stability, thus indicating the emergence of the Lys48-Asp73 electrostatic interaction might have been crucial for the evolution of SUMO domain folding, but not sufficient since further optimization of the position 48 environment might have been necessary to reach SUMO-1 stability. It is interesting to point out that the residue in position 74 in yeasts SUMO (smt3 and pmt3) is and Asp, while three consecutive Asp are found in positions 73 to 75 of *C. elegans* SUMO (Fig. 3). Moreover, the loop connecting β4 and α2/β5 adopts a different configuration in smt3 than in human SUMO-1/2; it is, therefore, tempting to hypothesize that reorganization within this loop, leading to a more geometrically favorable Asp arrangement, could have buffered the impact of emergence of a charged Lys in position 48.

Insights into the SUMO Folding Mechanism

The ubiquitin-like or β-grasp is one of the nine most frequent superfolds (Orengo et al. 1994), characterized by an ββαββ topology, and is populated by sequences sharing very low sequence homology (Bayer et al. 1998). Ubiquitin (Ub), apart from giving the name to this superfold, is one of the proteins whose folding pathway and mechanism has been more thoroughly studied (Jackson 2006), and thus provides an exceptional framework to analyze the folding of Ubiquitin-like proteins (Ubls). Although it was long discussed whether Ub folded through a two or three-state pathway (Went et al. 2004; Crespo et al. 2006; Khorasanizadeh et al. 1996; Krantz & Sosnick 2000), extensive protein engineering experiments showed that Ub folds through a highly polarized transition state (Went & Jackson 2005) without the presence of stable intermediates. This folding pathway for Ub was later corroborated by the first all-atom MD folding simulation in the ms timescale (Piana et al. 2013), which provides increased levels of molecular detail by showing how the first β-hairpin an part of the extended α-helix nucleate the Ub folding and dominate the structure of the TSE. This pattern of nucleation of the folding pathway based on the initial formation of a β-hairpin with a certain packing of the extended helix against it, has also been observed for other members of the Ub superfold like proteins G and L (McCallister et al. 2000; Kim et al. 2000). The ββα motif has been shown to dominate the transition states of many other α/β or α+β proteins too (Lindberg & Oliveberg 2007) and has subsequently been proposed as a foldon unit, that is, a minimal polypeptide sequence with ability to cooperatively fold by itself.

Since shifting the amino acids at positions 48 and 73 in the SUMO domain leads to a switch in its folding kinetics it can be subsequently postulated that interactions between residues occupying these positions must be strongly relevant for its folding mechanism. ϕ-analysis was carried at them by mutating the wt residues to Ala. While the E73A substitution in SUMO-2 has little impact and merely destabilizes slightly the unfolding kinetics, M48A has, in contrast, a ϕ-value ≈0.43 (Fig. 5.A) indicating that although not crucial for the formation of the TS, it may have a certain involvement. On the other hand, mutation to Ala in the equivalent positions of SUMO-1 results in an astonishing outcome. None of the mutants shows any variation of its Trp signal either at refolding or unfolding and no significant transition is observed for them when following their thermal unfolding by fluorescence or CD (Fig 5.B). While the fluorescence spectra of S1* is dominated by the signal at the Tyr emission maximum, the SUMO-1 Ala mutants present increased emission Trp accompanied by a red shift of the maximum intensity, which is more significant for the D73A variant, thus indicating the Trp probe is extensively exposed to solvent in these mutants to Ala, being its exposure slightly greater in D73A (Fig 5.C). CD spectrum shows K48A retains a significant content of secondary structure with an increase of the β-signature relative to the S1* spectra (Fig 5.D). In contrast, D73A exhibits a spectra characteristic of intrinsically unstructured proteins with negative signal around 200nm. It is interesting to note that D73A had to be analyzed at a lower concentration since it is produced as an insoluble protein and it tends to aggregate in the absence of denaturants.

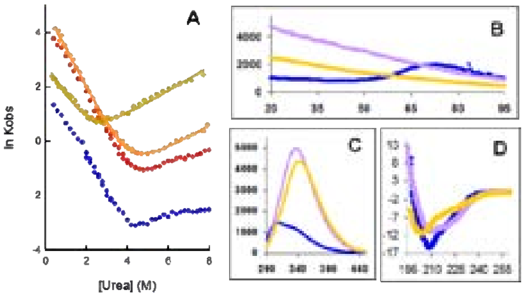


Figure. 5.- ϕ-analysis of positions 48 and 73 of SUMO Domain. A) Refolding and unfolding kinetics of S1*, in blue, S2*, in red and S2* variants M48A, in ocre, and E73A, in orange. B) thermal unfolding by intrinsic fluorescence, C) Intrinsic Trp fluoresece, and D) far-UV CD spectra of S1*, in blue, and its variants K48A, in magenta, and D73A in ocre.

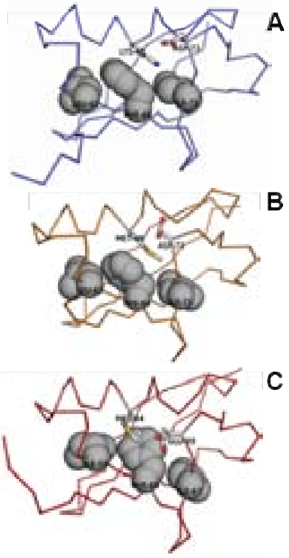


Figure 6.- Snapshot at ≈80ns of simulations showing the interactions of residue at position 48 with residues in β3/β4: A) wt SUMO-1, B) K48M, and C) wt SUMO-2.

The analysis of native state dynamics by MD of the environment at positions 48 and

73, reveals Met48 in SUMO-2 is mostly buried all along the simulation and contacts Ile62 and Phe64 (Fig 6). In SUMO-1, the Lys48 side chain aliphatic moiety is mostly buried as well, but instead makes contact with Phe64 in β3 and Ile71 in β4 although it can also slide through β3 to contact Leu62 and Phe64 (Fig 6.A). As noted before, Lys48 charged amino group establishes electrostatic interactions with the carboxyl group of Asp73 which has a vast freedom of rotation around its Cα, thus offering a large flexibility to maintain the electrostatic interaction with Lys ε-NH_2 as its hydrocarbon part undergoes configurational changes. When Lys48 is substituted by Met in the K48M mutant, the new side chain is accommodated between Phe64 and Ile71, as in the wt protein although it is buried deeper into the hydrophobic core due to the nonpolar nature of this thiomethyl terminus (Fig 6.B). However it is also observed to move towards the Nter of β3 and establish contacts with Leu62 and Phe64. The critical rate observed for position 48 in defining the folding kinetics of the SUMO domain and the close contacts it establishes with β3/β4 region particularly with Leu/Ile62, Phe64 and Ile71, suggest this region might be very relevant for the formation of SUMO TSE. Interestingly, the ϕ-values determined for the structurally equivalent positions in Ub β3/β4 are very low (Went & Jackson 2005), thus either the SUMO TSE is not polarized and extends all across the structure or it is polarized towards the opposite side, Cter, of the fold, as it has been observed in proteins G and L, which fold through TS with opposite polarization (McCallister et al. 2000; Kim et al. 2000). Whatever the case, packing of the extended α-helix against β3 or β3/β4 through burial of the amino acid at position 48 appears to be a crucial step for the formation of the TS in human SUMO domains. Collapse of the non-polar Met in SUMO-2 within the hydrophobic core would be favored by the establishment of interactions with hydrophobic residues in β3. However, burying a charged Lys comes at an increased cost and requires the action of a counter-ion able to exert some shielding of the charges. Following the folding and unfolding kinetics by pressure-jump provides additional clues to reveal how the SUMO TS may be formed. While SUMO2 possesses a regular activation volume profile along the folding coordinate (Fig. 7), which is indicative of a increased compaction of the polypeptide chain as it progresses towards the folded state, SUMO1 presents, conversely, a higher activation volume in the TS relative to both the folded and the unfolded state. Activation volumes calculated for the different conformations of the polypeptide can be rationalized in terms of the different structuration patterns of water molecules relative to the physico-chemical properties of the side chains they may interact with. Water molecules are considered to adopt a more ordered “lattice-like” structure around molecules with unfavorable interactions (hydrophobic) with them (Raschke & Levitt 2005; Sun et al. 2014). These ordered patterns of water imply a lower volume per ml of water molecules. In the unfolded state proteins exhibit a higher exposure to solvent of hydrophobic residues so water molecules would tend to form ordered structures around hydrophobic patches, resulting in a reduced volume of the solvation layer in the polypeptide chain. As the protein condenses as it advances towards the folded state, hydrophobic residues tend to hide in the core and ordered water structures are disrupted, subsequently increasing the volume of the solvation layer. In the case of the SUMO domains, paralogue 2 would proceed according to this scheme, but in SUMO-1 it has been already noted that burial of Lys48 is a limiting step in its folding. The formation of the TS will certainly imply the collapse of hydrophobic residues in the SUMO core, thus increasing the volume of the solvation layer, but, additionally, the effective burial of the Lys requires a certain conformational search of Lys48 and Asp73 side chains during the TS until the proper geometry is found for a stable interaction between their amino and carboxyl group. The transitory presence of these free charges in the TS as the conformational search takes place, would avoid unfavorable interactions between surrounding hydrophobic groups and water molecules. Upon consolidation of the Lys-Asp contact, the electrostatic potential of α-amino and γ-carboxyl groups is partially shielded and additional hydrophobic patches might be exposed leading to a decrease in the activation volume. This mechanism is consistent with both the kinetics of K48M and the conformational properties of the K48A and the D73A mutants. When the Lys is mutated to Ala, the substitution of the side chain with a methyl group allows a collapse of the chain that might be sufficient to attain a native-like topology and adopt certain native secondary structure preference, as revealed by the CD spectrum, in a molten-globule state which, however, is not capable of condensing completely to achieve the folded state - lacking necessary interactions at position 48. When the aliphatic chain is extended by means of the Met side chain, SUMO is able to fold properly but through a different pathway which avoids the conformational search required for burying a Lys, likely through TS that is more similar to that of SUMO-2. CD spectrum of the D73A mutant indicates that suppression of the counter-ion in position 73 completely abrogates the capability of Lys48 to be buried. Impeding the aliphatic part of its side chain to participate in the interactions that nucleate the folding of the protein, thus no native-like topology can be attained and the polypeptide remains unstructured. As can be derived from its folding constant, SUMO-1 pathway proceeds through a higher folding barrier relative to that in the SUMO-2 pathway, which is likely due to the energetic penalty of Lys burial. In return, SUMO-1 possesses a larger unfolding barrier too, which may have certain functional implications.

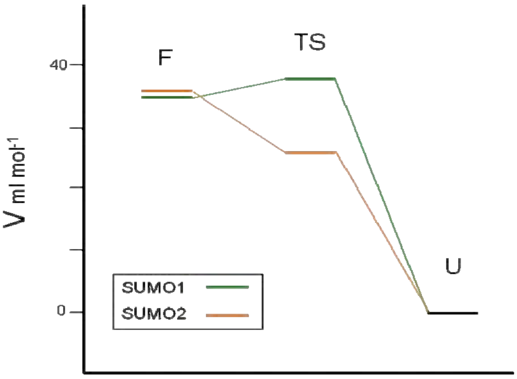


Figure 7.- Activation Volumes calculated for wt SUMO1 and wt SUMO2 in the unfolded (U), the folded (F) and the transition states.

Biological Significance of Divergently Evolved Folding Pathways in SUMO Domains

The presence in vertebrates of two main forms of a protein which does not possess a function *per se*, raises the question about the specific activity of the SUMO domains. Although SUMOylation of the same substrate by one or the other form has been observed to result in different functional outcomes (Wilkinson & Henley 2010), the determinants accounting for the differential SUMO specificity at the different layers of the modification pathway (processing, substrate conjugation and deconjugation) are not completely understood yet, particularly in humans (Flotho & Melchior 2013; Gareau & Lima 2010); SUMO proteases and ligases may present either similar or very dissimilar affinity for the different forms and while some proteins may only be covalently modified with one of the SUMO domains, many can be equally modified with any of them (Vertegaal et al. 2006).

In order to shed light as to whether swapping the folding kinetics of SUMO domains may have any impact on SUMOylation, the conjugation of RanGAP1 (the first SUMO substrate to be identified (Matunis 1996; Mahajan et al. 1997)) with either wt or engineered SUMO domains was evaluated (Fig. 8). Reaction results show the engineered proteins are fully functional and SUMOylate RanGAP1 until exhaustion of the substrate. Moreover, conjugation rates are undistinguishable for full-length wt SUMO-1 and S1*K48M, and very similar for wt SUMO-2 (employing the variant lacking the N_{ter} tail as described above -also noted SUMO-2Δ14- to avoid self-conjugation) and S2*M48K+E73D, thus suggesting that mutations driving changes in the SUMO folding pathway do not compromise its functionality, neither are translated into any shift in SUMO specificity.

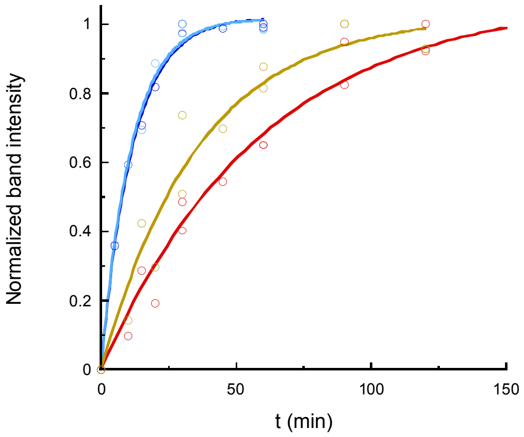


Fig. 6.- Functional Analysis of SUMO Domains Variants. Kinetics of the RanGAP1 SUMOylation with SUMO-1wt (dark blue), S1*K48M (light blue), SUMO-2Δ14 (red) and S2*M48K+E73D (ocre)

A putative differential conformational flexibility, arising from the distinctive unfolding kinetics of the SUMO domains, does not seem likely to possess any relevance in the process of SUMO conjugation since the mechanisms of SUMO activation by proteases, and isopeptide bond catalysis by transferases and ligases merely imply the action of these enzymes in the environment of the diglycine motif, which is exposed at the unstructured C_{ter} tail. Thus no conformational rearrangement is required to perform these steps. Aswell, the interactions established between the SUMO proteins and the enzymes participating of the modification pathway, and even with SUMO substrates, do not appear to require noticeable conformational changes in SUMO, provided the modifier structure remains broadly unaltered in the many different complexes which have been structurally solved to date (Geiss-Friedlander & Melchior 2007).

Such plausible differences in SUMO dynamic flexibility neither seem to be relevant for those recognition modes within the SUMOylation pathway governed by protein complex formation through SUMO interacting motifs (SIMs); taking into account these interactions are basically ruled by the primary structure and surface charge distribution (Hecker et al. 2006; Namanja et al. 2012). Furthermore, SUMO domains interact with SIMs through an exposed hydrophobic patch on its surface (Sabate et al. 2012), so no significant conformational reconfiguration is required to ensure the complex formation.

A different source of regulation of paralogue specificity in SUMOylation appears to be exerted at the level of modifier deconjugation. For example, RanGAP1 preferential specificity towards SUMO-1 has been shown to arise from the ability of RanGAP1 modified with paralogue 1 to form a more stable interaction with its Ubc9 and RanBP2 partners, than that established by RanGAP1~SUMO-2, thus *in vivo* steady-state levels of the first are higher (Zhu et al. 2009). Interestingly, SUMO-1 is mostly found conjugated to substrates in mammalian cells, whereas paralogue 2/3 is mainly found on a free form (Saitoh & Hinchey 2000), possibly due to a greater SUMO-2/3 specific deconjugase activity in the cell (Kolli et al. 2010); this yields a pool of free SUMO 2/3 which has been proposed to serve as a reservoir. In fact, increased levels of substrate conjugation with SUMO-2/3 have been reported in response to the induction of different cellular stresses (Wilkinson & Henley 2010), although the specificity of SUMO modification may also depend on the intensity and duration of the induced stress (Tempé et al. 2008).

Nonetheless, in the case the maintenance of a free SUMO-2/3 pool in the physiologically averaged steady-state of the cell is certainly dependent on a higher paralogue 2/3 specific deconjugase activity, then an increased conformational flexibility stemming from higher unfolding rates in SUMO-2, relative to those of SUMO-1, might actually have a certain engagement in the ease of modifier deconjugation, since the SUMO-substrate bond needs to dock to the protease active site, which involves the isomerization of the isopeptidic bond (Reverter & Lima 2006; Shen et al. 2006).

Unfolding rates are related to the height of the activation barrier in the unfolding side of the reaction coordinate, which allows to estimate proteins effective stability in the cellular environment or kinetic stability. Kinetic stability possesses a prominent biological significance since it influences relevant processes like the population of unfolded or partially unfolded states or resistance to proteolytic degradation (Sanchez-Ruiz 2010). Considering the majority of SUMO-2/3 is present as unconjugated modifier in the steady-state, it may be subject to a high turnover rate in the cell. Faster unfolding kinetics of SUMO-

2 could be translated into a lower kinetic stability, relative to SUMO-1, which might favor this turnover by making paralogue 2 more prone to be degraded by the cellular proteolytic machinery.

8.3.- References

Bayer, P. et al., 1998. Structure determination of the small ubiquitin-related modifier SUMO-1. *Journal of molecular biology*, 280(2), pp.275–86.

Crespo, M.D., Simpson, E.R. & Searle, M.S., 2006. Population of on-pathway intermediates in the folding of ubiquitin. *Journal of molecular biology*, 360(5), pp.1053–66.

Dalessio, P.M. et al., 2005. Swapping core residues in homologous proteins swaps folding mechanism. *Biochemistry*, 44(8), pp.3082–90.

Flotho, A. & Melchior, F., 2013. Sumoylation: a regulatory protein modification in health and disease. *Annual review of biochemistry*, 82, pp.357–85.

Friel, C.T., Beddard, G.S. & Radford, S.E., 2004. Switching two-state to three-state kinetics in the helical protein Im9 via the optimisation of stabilising non-native interactions by design. *Journal of molecular biology*, 342(1), pp.261–73.

Gareau, J.R. & Lima, C.D., 2010. The SUMO pathway: emerging mechanisms that shape specificity, conjugation and recognition. *Nature reviews. Molecular cell biology*, 11(12), pp.861–71.

Geiss-Friedlander, R. & Melchior, F., 2007. Concepts in sumoylation: a decade on. *Nature reviews. Molecular cell biology*, 8(12), pp.947–56.

Graña-Montes, R. et al., 2014. N-terminal protein tails act as aggregation protective entropic bristles: the SUMO case. *Biomacromolecules*, 15(4), pp.1194–203.

Hay, R.T., 2005. SUMO: a history of modification. *Molecular cell*, 18(1), pp.1–12.

Hecker, C.-M. et al., 2006. Specification of SUMO1- and SUMO2-interacting motifs. *The Journal of biological chemistry*, 281(23), pp.16117–27.

Jackson, S.E., 2006. Ubiquitin: a small protein folding paradigm. *Organic & biomolecular chemistry*, 4(10), pp.1845–53.

Johnson, E.S., 2004. Protein modification by SUMO. *Annual review of biochemistry*, 73, pp.355–82.

Khorasanizadeh, S. et al., 1993. Folding and stability of a tryptophan-containing mutant of ubiquitin. *Biochemistry*, 32(27), pp.7054–7063.

Khorasanizadeh, S., Peters, I.D. & Roder, H., 1996. Evidence for a three-state model of protein folding from kinetic analysis of ubiquitin variants with altered core residues. *Nature Structural Biology*, 3(2), pp.193–205.

Kim, D.E., Fisher, C. & Baker, D., 2000. A breakdown of symmetry in the folding transition state of protein L. *Journal of molecular biology*, 298(5), pp.971–84.

Kolli, N. et al., 2010. Distribution and paralogue specificity of mammalian deSUMOylating enzymes. *The Biochemical journal*, 430, pp.335–344.

Kotamarthi, H.C., Sharma, R. & Koti Ainavarapu, S.R., 2013. Single-molecule studies on PolySUMO proteins reveal their mechanical flexibility. *Biophysical journal*, 104(10), pp.2273–81.

Krantz, B. a & Sosnick, T.R., 2000. Distinguishing between two-state and three-state models for ubiquitin folding. *Biochemistry*, 39(38), pp.11696–701.

Kumar, A. et al., 2006. Residue-level NMR view of the urea-driven equilibrium folding transition of SUMO-1 (1-97): native preferences do not increase monotonously. *Journal of molecular biology*, 361(1), pp.180–94.

Lappalainen, I., Hurley, M.G. & Clarke, J., 2008. Plasticity within the obligatory folding nucleus of an immunoglobulin-like domain. *Journal of molecular biology*, 375(2), pp.547–59.

Larson, S.M. et al., 2002. Residues participating in the protein folding nucleus do not exhibit preferential evolutionary conservation. *Journal of molecular biology*, 316(2), pp.225–33.

Lawrence, C. et al., 2010. Investigation of an anomalously accelerating substitution in the folding of a prototypical two-state protein. *Journal of molecular biology*, 403(3), pp.446–58.

Lindberg, M.O. & Oliveberg, M., 2007. Malleability of protein folding pathways: a simple reason for complex behaviour. *Current opinion in structural biology*, 17(1), pp.21–9.

Lowe, A.R. & Itzhaki, L.S., 2007. Rational redesign of the folding pathway of a modular protein. *Proceedings of the National Academy of Sciences of the United States of America*, 104(8), pp.2679–84.

Mahajan, R. et al., 1997. A small ubiquitin-related polypeptide involved in targeting RanGAP1 to nuclear pore complex protein RanBP2. *Cell*, 88, pp.97–107.

Matunis, M.J., 1996. A novel ubiquitin-like modification modulates the partitioning of the Ran-GTPase-activating protein RanGAP1 between the cytosol and the nuclear pore complex. *The Journal of Cell Biology*, 135(6), pp.1457–1470.

McCallister, E.L., Alm, E. & Baker, D., 2000. Critical role of beta-hairpin formation in protein G folding. *Nature structural biology*, 7(8), pp.669–73.

Mirny, L. & Shakhnovich, E., 2001. Evolutionary conservation of the folding nucleus. *Journal of molecular biology*, 308(2), pp.123–9.

Morcos, F. et al., 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America*, 108(49), pp.E1293–301.

Namanja, A.T. et al., 2012. Insights into high affinity small ubiquitin-like modifier (SUMO) recognition by SUMO-interacting motifs (SIMs) revealed by a combination of NMR and peptide array analysis. *The Journal of biological chemistry*, 287(5), pp.3231–40.

Nauli, S., Kuhlman, B. & Baker, D., 2001. Computer-based redesign of a protein folding pathway. *Nature structural biology*, 8(7), pp.602–5.

Nickson, A. a & Clarke, J., 2010. What lessons can be learned from studying the folding of homologous proteins? *Methods (San Diego, Calif.)*, 52(1), pp.38–50.

Nickson, A. a, Wensley, B.G. & Clarke, J., 2013. Take home lessons from studies of related proteins. *Current opinion in structural biology*, 23(1), pp.66–74.

Oliveberg, M., 1998. Alternative Explanations for “Multistate” Kinetics in Protein Folding: Transient Aggregation and Changing Transition-State Ensembles †. *Accounts of Chemical Research*, 31(11), pp.765–772.

Orengo, C. a, Jones, D.T. & Thornton, J.M., 1994. Protein superfamilies and domain superfolds. *Nature*, 372(6507), pp.631–4.

Owerbach, D. et al., 2005. A proline-90 residue unique to SUMO-4 prevents maturation and sumoylation. *Biochemical and biophysical research communications*, 337(2), pp.517–20.

Piana, S., Lindorff-Larsen, K. & Shaw, D.E., 2013. Atomic-level description of ubiquitin folding. *Proceedings of the National Academy of Sciences of the United States of America*, 110(15), pp.5915–20.

Raschke, T.M. & Levitt, M., 2005. Nonpolar solutes enhance water structure within hydration shells while reducing interactions between them. *Proceedings of the National Academy of Sciences of the United States of America*, 102(19), pp.6777–82.

Reverter, D. & Lima, C.D., 2006. Structural basis for SENP2 protease interactions with SUMO precursors and conjugated substrates. *Nature structural & molecular biology*, 13, pp.1060–1068.

Sabate, R. et al., 2012. Native structure protects SUMO proteins from aggregation into amyloid fibrils. *Biomacromolecules*, 13(6), pp.1916–26.

Saitoh, H. & Hinchey, J., 2000. Functional Heterogeneity of Small Ubiquitin-related Protein Modifiers SUMO-1 versus SUMO-2/3. *Journal of Biological Chemistry*, 275(9), pp.6252–6258.

Sanchez-Ruiz, J.M., 2010. Protein kinetic stability. *Biophysical chemistry*, 148(1-3), pp.1–15.

Shen, L. et al., 2006. SUMO protease SENP1 induces isomerization of the scissile peptide bond. *Nature structural & molecular biology*, 13, pp.1069–1077.

Sun, T. et al., 2014. An antifreeze protein folds with an interior network of more than 400 semi-clathrate waters. *Science (New York, N.Y.)*, 343(6172), pp.795–8.

Tempé, D., Piechaczyk, M. & Bossis, G., 2008. SUMO under stress. *Biochemical Society transactions*, 36(Pt 5), pp.874–8.

Valdar, W.S.J., 2002. Scoring residue conservation. *Proteins*, 48(2), pp.227–41.

Vertegaal, A.C.O. et al., 2006. Distinct and overlapping sets of SUMO-1 and SUMO-2 target proteins revealed by quantitative proteomics. *Molecular & cellular proteomics□: MCP*, 5, pp.2298–2310.

Watters, A.L. et al., 2007. The highly cooperative folding of small naturally occurring proteins is likely the result of natural selection. *Cell*, 128(3), pp.613–24.

Went, H.M., Benitez-Cardoza, C.G. & Jackson, S.E., 2004. Is an intermediate state populated on the folding pathway of ubiquitin? *FEBS letters*, 567(2-3), pp.333–8.

Went, H.M. & Jackson, S.E., 2005. Ubiquitin folds through a highly polarized transition state. *Protein engineering, design & selection□: PEDS*, 18(5), pp.229–37.

Wilkinson, K. a & Henley, J.M., 2010. Mechanisms, regulation and consequences of protein SUMOylation. *The Biochemical journal*, 428(2), pp.133–45.

Zhu, S. et al., 2009. Protection from Isopeptidase-Mediated Deconjugation Regulates Paralog-Selective Sumoylation of RanGAP1. *Molecular Cell*, 33, pp.570–580.