CHAPTER THREE

Physicochemical Principles of Protein Aggregation

Benedetta Bolognesi, Gian Gaetano Tartaglia
Centre for Genomic Regulation, CRG and UPF, Barcelona, Spain

Contents

Abstract

This chapter provides a theoretical framework on the quantitative description of protein aggregation. The reader is provided with an overview of the fundamental theory of linear and helical polymers, as well as an introduction on the parameters governing evolution of aggregates over time. The models presented for the interpretation of the protein aggregation process take into account the contributions of different physicochemical parameters such as charge, hydrophobicity, and secondary structure propensity. Finally, we discuss our current understanding of how prediction of aggregation rates and identification of aggregation-prone protein regions are predicted from the information contained in the primary amino acid sequence.

1. INTRODUCTION

β-Sheet structures are, together with α helices, the most common regular motifs in natively folded proteins. Partial or complete disruption of the native fold is observed when a protein is subjected to stress originating from

Progress in Molecular Biology and Translational Science, Volume 117
ISSN 1877-1173
http://dx.doi.org/10.1016/B978-0-12-386931-9.00003-9

unsuitable chemical (e.g., high or low pH, high salt concentrations, hydrophobic environment) or physical (e.g., high temperature, high pressure) agents.[1,2] Denatured proteins have no defined secondary and tertiary structure and, especially at high concentrations, tend to aggregate into insoluble deposits. Many aggregates, known as amyloid fibrils, share a common morphology and can be recognized under the electron microscope as regular rope-like structures that measure micrometers in length and a few nanometers in diameter. As shown by biophysical techniques, such as circular dichroism and Fourier transform infrared spectroscopy, amyloid fibrils have a high content of β structure, whichever the structure of the monomeric molecule in the native state. X-Ray diffraction analysis of fibrils yields a typical cross-β diffraction pattern, signature of an intermolecular β-sheet structure, where the hydrogen bonding among β strands runs parallel to the main fibril axis generating a pleated β-sheet structure. Investigations using electron and atomic force microscopy show that amyloid fibrils consist of a variable number of protofilaments, each of a diameter of approximately two nanometers and twisted around each other to form supercoiled rope-like structure.[3,4] Packing of the filaments is dependent on the protein system, but a single system can also display multiple fibril morphologies.[5]

One of the main causes of incorrect protein folding *in vivo* is cell stress, which can be caused by heat shock, nutrient depletion, or other stimuli.[6,7] Production of inactive proteins not only represents an energetic drain and a metabolic load for the cell but also may result in accumulation of the unfolded proteins within inclusion bodies that are responsible for cell damage. Indeed, misfolded proteins that escape the quality control mechanisms of the cell may lead to the impairment of relevant biological processes and affect the viability of the organism. Up to now, protein aggregation has been associated with more than 30 diseases and in particular amyloid fibrils have been found involved in a number of debilitating pathologies including Alzheimer's, Parkinson's, Huntington's, prion disease, and type II diabetes.[8,9] The propensity of different proteins to form amyloid fibrils can vary widely depending on the physicochemical properties of the specific amino acid sequence involved.[10,11]

Misfolding and aggregation of proteins *in vivo* differ from similar processes taking place under *in vitro* experimental conditions as they occur in complex cellular environments containing a host of factors that are known to modulate protein aggregation and attempt to protect against any subsequent toxicity.[12] In fact, efficient folding of many newly synthesized proteins depends on assistance from molecular chaperones, which prevent protein misfolding and aggregation in the crowded environment of the cell.[13] It is estimated that more than 30% of the newly synthesized proteins are degraded by proteasome due to translation errors or improper folding.[14] Nevertheless, it is important to mention that the same physicochemical principles that lead to aggregation *in vitro* are present in cellular environments, and once the physicochemical determinants for aggregation are identified, they can be conveyed almost straightforwardly into a mathematical model.

In this chapter, we discuss physicochemical principles that lie behind protein polymerization, including kinetic models for oligomerization and evolutionary pressures arising against toxic aggregation.

2. LINEAR POLYMERS

Let us consider a solution of macromolecules with the ability to form aggregates by end-to-end association (Fig. 3.1A). In equilibrium state, the solution contains dispersed monomers and linear polymers of various lengths. The following mass action law[15] exists between the concentration of monomers μ_1 and dimers μ_2

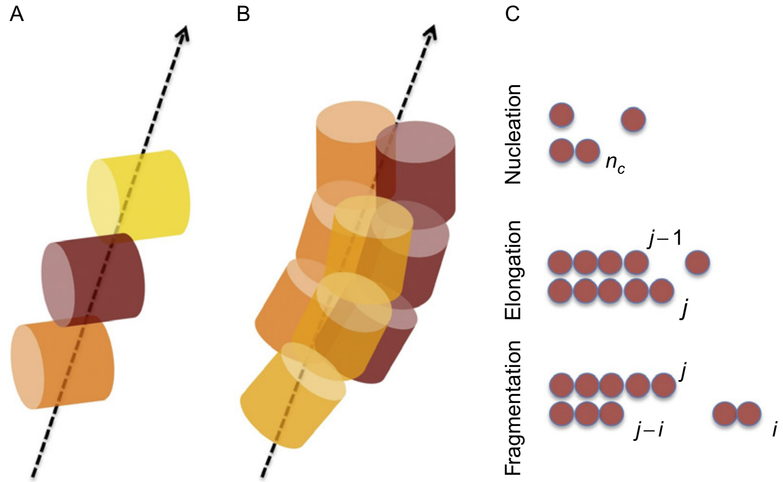

Figure 3.1 Protein oligomerization. Models of protein aggregation: (A) linear polymers and (B) helical polymers. In amyloid fibrils, each building block will acquire a β structure once incorporated in the polymer. (C) Growth trough nucleation, elongation, and fragmentation.

$$\mu_2 = z_1 \mu_1^2 \qquad [3.1]$$

where z_1 is the equilibrium constant of dimerization. Similarly for the transition from $(i+1)$-mer to (i)-mer

$$\mu_{i-1} = z_i \mu_i \mu_1 \qquad [3.2]$$

When z_1 is independent of i or when the binding free energy of monomer to i-mer is independent of i, the equilibrium concentration of the i-mer is given by

$$\mu_i = z^{-1}(z\mu)^i \qquad [3.3]$$

In Eq. (3.3), we have a function of the equilibrium constant $z(=z_i)$ and monomer concentration $\mu(=\mu_i)$.

The total mass m can be expressed as

$$m = \sum_{i=1} i\mu_i \qquad [3.4]$$

If z_i is independent of i, we have

$$m = \sum iz^{-1}(z\mu)^i = z^{-1}\frac{z\mu}{(1-z\mu)^2} \qquad [3.5]$$

Using Eq. (3.5), we can determine the concentration of monomer at given m, and from μ we can calculate the concentration of i-mers.

In general, the following scheme can be derived when the aggregates are in equilibrium with soluble monomeric material $M_S = \mu_1$

$$M_S + \mu_{i-1} \leftrightarrow \mu_i \qquad [3.6]$$

where the equilibrium constant is $Z_{eq} = \mu_i / M_S \mu_{i-1}$.

In this case, the total mass concentration $m = \sum_{i=1} i\mu_i = M_S/(1 - Z_{eq}M_S)^2$ can be rearranged as

$$Z_{eq} = \frac{1}{M_S} - \frac{1}{\sqrt{M_S m}} \qquad [3.7]$$

When $M_S \gg m$, the equilibrium expression reduces to the situation where fibril ends E are in equilibrium with monomers

$$M_S + E \leftrightarrow E \qquad [3.8]$$

where the equilibrium constant is $Z_{eq} = 1/M_S$.

3. HELICAL POLYMERS

In the helix formed by a linear polypeptide, each amino acid is bound to ~4 amino acids by two kinds of bonds, one being the primary bond with neighboring amino acids in the polypeptide chain and the other the hydrogen bond with the third preceding amino acid along the chain. Similarly, aggregates can be formed by monomer macromolecules with two bonds with other monomers. In the linear polymer, each monomer is bound to two other monomers and usually all bonds are of the same chemical nature (Fig. 3.1A). In the helical polymer having three monomers per turn, each monomer is bound with neighboring monomers along the linear chain and simultaneously with the third preceding and succeeding monomers (Fig. 3.1B). If we consider an equilibrium solution containing both linear polymers and helical polymers,[15] the concentration of the helical trimer $\mu_{3\mathrm{h}}$ can be related to that of the linear trimer $\mu_{3\mathrm{l}}$ by

$$\mu_{3\mathrm{h}} = \sigma\mu_{3\mathrm{l}} \qquad [3.9]$$

where $\sigma = \exp(-\partial f/KT)$ and ∂f is the free energy increment for helical trimers.

If we introduce the chemical constant z_h of equilibrium between the fourth monomer and the helical trimer, the concentration of the shortest helical polymer $\mu_{4\mathrm{h}}$ will follow

$$\mu_{4\mathrm{h}}/\sigma = z_\mathrm{h}\mu_{3\mathrm{h}}\mu_1/\sigma = \mu z^{-1} z_\mathrm{h}(z\mu)^3 = z^{-1}(z/z_\mathrm{h})^3(z_\mathrm{h}\mu)^4 \qquad [3.10]$$

Since the fourth monomer in the helix can bind two monomers, z_h is usually larger than z. For the further growth of the helical polymer by attaching monomers to the helical nucleus, we can assume the same chemical constant z_h and obtain

$$\mu_{i\mathrm{h}} = z^{-1}(z/z_\mathrm{h})^3(z_\mathrm{h}\mu)^i\sigma = z^{-1}\nu(z_\mathrm{h}\mu)^i \qquad [3.11]$$

where $\nu = \sigma(z/z_\mathrm{h})^3$.

In the solution containing monomers and linear and helical polymers, we have

$$m = \mu + m_\mathrm{l} + m_\mathrm{h} \qquad [3.12]$$

where $m_\mathrm{l} = \sum_{i=2} i z^{-1}(z\mu)^i$ and $m_2 = \sum_{i=3} i z^{-1}\nu(z_\mathrm{h}\mu)^i$.

For very small values of m, monomer concentration μ increases proportionally to m and a small number of linear polymers (dimers, trimers, etc.)

will appear in solution. When μ approaches z^{-1}, helical polymers will start to appear (if z_h is larger than z) and the total concentration approaches

$$m_\mathrm{c} = \frac{1/z_\mathrm{h}}{(1 - z/z_\mathrm{h})^2} \qquad [3.13]$$

If $m \gg m_\mathrm{c}$, $m \sim m_\mathrm{h}$ and helical polymers prevail in solution.[15]

4. TIME EVOLUTION OF LINEAR AND HELICAL POLYMERS

Let us denote the concentration of monomers, linear (i)-mers, and helical (i)-mers at time t as $\mu(t)$, $\mu_{i\mathrm{l}}(t)$, and $\mu_{i\mathrm{h}}(t)$, respectively.[15] The growth rate of helical (i)-mers to helical $(i+1)$-mers can be expressed as $k_+\mu(t)\mu_{i\mathrm{h}}(t)$ and the rate of detachment from (i)-mers will be $k_-\mu_{i\mathrm{h}}(t)$. Similarly, the rate of transformation from linear to helical trimers can be expressed as $k'_+\mu_{3\mathrm{l}}(t)$ and the reverse reaction will be driven by $k'_-\mu_{3\mathrm{h}}(t)$. Hence, we have for the total number concentration[16]

$$\frac{\mathrm{d}p}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\left[\sum_{i=3}\mu_{i\mathrm{h}}(t)\right] = k'_+\mu_{3\mathrm{l}}(t) - k'_-\mu_{3\mathrm{h}}(t) \qquad [3.14]$$

and for the mass concentration

$$\begin{aligned}\frac{\mathrm{d}m}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\left[\sum_{i=3} i\mu_{i\mathrm{h}}(t)\right] &= \left[\sum_{i=3}(k_+\mu(t) - k_-)\mu_{i\mathrm{h}}(t)\right] \\ &\quad + 3[k'_+\mu_{3\mathrm{l}}(t) - k'_-\mu_{3\mathrm{h}}(t)] + k_-\mu_{3\mathrm{h}}(t) \qquad [3.15]\end{aligned}$$

The increasing rate of the total mass m_h participating in helical polymers or the decreasing rate of $\mu + m_\mathrm{l}$ of monomers and linear polymers can be calculated as

$$\begin{aligned}-[\mathrm{d}(\mu + m_\mathrm{l})/\mathrm{d}t] = \mathrm{d}m_\mathrm{h}/\mathrm{d}t &= [k_+\mu(t) - k_-]\int \mathrm{d}t\,[k'_+\mu_{3\mathrm{l}}(t) - k'_-\mu_{3\mathrm{h}}(t)] \\ &\quad + 3[k'_+\mu_{3\mathrm{l}}(t) - k'_-\mu_{3\mathrm{h}}(t)] \\ &\quad + k_-\mu_{3\mathrm{h}}(t)\end{aligned} \qquad [3.16]$$

Assuming that $\mu_{3\mathrm{l}}$ and $\mu_{3\mathrm{h}}$ are proportional to μ^3 (the polymerization–depolymerization reaction is more rapid than helix formation), we have for $m_\mathrm{l} \ll \mu$

$$-\mathrm{d}\mu/\mathrm{d}t = (k_+\mu - k_-)\int c\mu^3 \mathrm{d}t \qquad [3.17]$$

where c is a constant.

Equation (3.17) shows how free monomer concentration decreases in time during the aggregation process.

When $k_+\mu \gg k_-$, the differential equation can be solved

$$\ln\frac{\left[1+(1-x^3)^{1/2}\right]\left[1-(1-x_0{}^3)^{1/2}\right]}{\left[1-(1-x^3)^{1/2}\right]\left[1+(1-x_0{}^3)^{1/2}\right]} = 3\alpha t \qquad [3.18]$$

where $x = x_0^3(\mu/\mu_0)^3$, $\alpha = \gamma^{1/2}x_0{}^{-3/2}$, $\gamma = (2/3)k_+c\mu_0^3$, $x_0{}^{-3} = 1 + k_+^3h_0^2/\gamma$, and μ_0 is the initial concentration of monomers. If the initial concentration of helical nuclei h_0 is negligible, we have

$$\ln\frac{\left[1+\left(1-\mu^3/\mu_0^3\right)^{1/2}\right]}{\left[1-(1-\mu^3/\mu_0^3)^{1/2}\right]} \approx \ln\left[4\mu_0^3/\mu^3 - 1\right] = 3\alpha t \qquad [3.19]$$

As indicated by Eq. (3.19), the concentration of free monomer μ decreases exponentially as a power law of μ_0.

5. TIME EVOLUTION OF FIBRILS

A master equation can be used to describe the time evolution of the concentration $\mu(t,j)$ of aggregates of length j[17]

$$\begin{aligned}\frac{\partial\mu(t,j)}{\partial t} &= 2m(t)k_+\mu(t,j-1) - 2m(t)k_+\mu(t,j)\\ &\quad + k_-(j-1)\mu(t,j) + 2k_-\sum_{i=j+1}\mu(t,i) + k_n m(t)^{n_c}\delta_{j,n_c}\end{aligned} \qquad [3.20]$$

where $m(t)$ is the concentration of monomers. The first term in Eq. (3.20) accounts for the increase in the number of filaments of length j due to the addition of monomers of either end of filament of length $j-1$ (Fig. 3.1C). The term $2m(t)k_+\mu(t, j)$ describes the decrease in the number of filaments of length j growing further to length $j+1$, while $k_-(j-1)\mu(t, j)$ reflects the possibility of a filament of length j breaking at any of its $j-1$ internal links. The term $2k_-\sum_{i=j+1}\mu(t,i)$ accounts for the fact that there are two links in any filament of length $i > j$ where breakage leads to a filament of length j, while

$k_n m(t)^{n_c} \delta_{j,n_c}$ represents the spontaneous formation of growth nuclei of size n_c (Fig. 3.1C). Hence, we have for the total number concentration

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t} = k_-[M(t) - (2n_c - 1)P(t)] + k_n m(t)^{n_c} \qquad [3.21]$$

and for the total mass concentration

$$\frac{\mathrm{d}M(t)}{\mathrm{d}t} = 2[m(t)k_+ - n_c(n_c - 1)k_-/2]P(t) + n_c k_n m(t)^{n_c} \qquad [3.22]$$

Using fixed-point analysis, the system can be integrated as

$$\begin{aligned} P(t) = \frac{m_{\mathrm{tot}}}{2n_c - 1} - \frac{m_{\mathrm{tot}} k_- \exp[-(-2n_c - 1)k_- t]}{\kappa} \\ Ei(-C_+ \mathrm{e}^{\kappa t}) + \exp[(-2n_c - 1)k_- t]B_2 \end{aligned} \qquad [3.23]$$

and

$$M(t) = m_{\mathrm{tot}}\left(1 - \exp\left[-C_+ \exp(\kappa t) + C_- \exp(-\kappa t) + k_n m_{\mathrm{tot}}^{n_c - 1} k_-^{-1}\right]\right) \qquad [3.24]$$

where $\kappa = \sqrt{2m_{\mathrm{tot}} k_+ k_-}$ is the rate of multiplication of filament population, $C_\pm = k_+ P(0)/\kappa \pm M(0)/(2m_{\mathrm{tot}}) \pm \left(k_n m_{\mathrm{tot}}^{n_c - 1}\right)/(2k_-)$ and $m_{\mathrm{tot}} = M(t) + m(t)$. Considering the steepest slope of the kinetic trace, we have $v_{\max} = \kappa/\log(1/C_+)$.

The variable $\kappa = \sqrt{2m_{\mathrm{tot}} k_+ k_-}$ defines the lag phase, which exists only if the growth rate $m_{\mathrm{tot}}\kappa/\mathrm{e}$ is maximal at $t_{\max} = \kappa^{-1} \log(1/C_+)$. More generally, the parameter κ, which corresponds to the rate of multiplication of the population of fragments, emerges as the most important quantity describing the overall properties of systems that self-assemble by processes that involve elongation and fragmentation. In a regime where secondary nucleation through fragmentation of filaments is an effective source of filaments than primary nucleation $\left(k_-/\left(k_n m_{\mathrm{tot}}^{n_c - 1}\right) \gg 1\right)$, observables such as the lag time and maximal growth rate depend primarily on just the single parameter κ[17] (Fig. 3.2A).

Very recently, a systematic investigation of bulk experimental measurements has highlighted the relevance of secondary pathways, other than fragmentation, in driving the overall aggregation reaction.[18] The analysis reveals a crucial role to existing aggregates, which would be able to accelerate the production of further aggregates, resulting in positive feedback type of mechanism. The approach, based on the standard model of filamentous growth first presented by Oosawa in the 1960s[16] and extended by Eaton

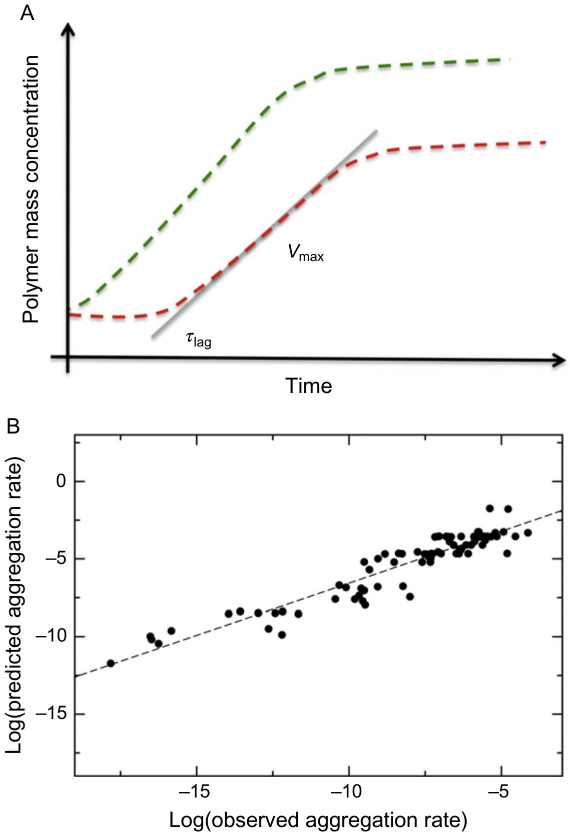

Figure 3.2 Aggregation kinetics. (A) Models for linear and helical polymerization (Eq. 3.18) can reproduce v_{max} in the exponential phase (green curve); equations based on nucleation, elongation, and fragmentation (Eq. 3.21) are used to describe lag phases (red curve). (B) The aggregation rates v_{max} can be accurately predicted using phenomenological formulas (Eq. 3.34).

in the 1980s,[19] provides a new framework which allows derivation of closed form analytical solutions for the lower principal moments of the fibril length distribution, as well as a range of accompanying scale laws.

6. THE AGGREGATION RATE

Beaven *et al.* have shown, already in 1969,[20] that under a critical concentration, the process of aggregation of glucagon is slow and that polymerization occurs more readily at high concentrations. More recently, Ruschak and Miranker have reported for the islet amyloid peptide that the rate of fibril elongation increases with monomer concentration, being the slope for the fitting very close to 1.[21] These findings are in good agreement with the fact that the rate of multiplication of filament population (Section 5) is proportional to the square of protein (monomer and fibril) in solution. In general, concentration, temperature, ionic strength, and pH are essential factors influencing the process of aggregation and can be regarded as *extrinsic contributions* to distinguish them from the *intrinsic contributions* that are inherent properties of the polypeptide chain, dependent on the amino acid position. With good approximation, the aggregation rate can be assumed to increase with temperature and concentration because the probability of collision and elongation of polypeptide chains increases with temperature and concentration. Although aggregation rate and temperature are not expected to correlate above physiological values,[22] the use of linear dependences is preferable for the small extent of experimental accessible values. In agreement with quasi-elastic light-scattering experiments of fibrillogenesis of the amyloid-β protein, the aggregation rate could be assumed to be proportional to the concentration c for Ref. 23 and to be independent of the concentration above the critical value $c=c^*$.[24]

7. INTRINSIC DETERMINANTS OF PROTEIN AGGREGATION

A significant correlation was reported between the changes in the aggregation rates resulting from single amino acid mutations and their effect on physicochemical properties such as hydrophobicity, charge, and the propensity to adopt α-helical or β-sheet secondary structures.[25] The different factors were included in an equation that predicts changes in aggregation rates relative to the wild-type protein under denaturing conditions[25]

$$\log(v_{\mathrm{wt}}/v_{\mathrm{mut}}) = \alpha_{\mathrm{hydr}}\Delta I^{\mathrm{hydr}} + \alpha_{\mathrm{ss}}\Delta I^{\mathrm{ss}} + \alpha_{\mathrm{ch}}\Delta I^{\mathrm{ch}} \qquad [3.25]$$

In this equation, $\log(v_{\mathrm{wt}}/v_{\mathrm{mut}})$ represents the logarithm of the ratio between v_{wt} and v_{mut}, the aggregation rates of wild-type and mutant sequence, respectively, and ΔI^{hydr}, ΔI^{ss}, and ΔI^{ch} represent the change in hydrophobicity, I^{hydr}, secondary structure propensity, I^{ss}, and electrostatic charge, I^{ch}, upon mutation. The parameters α were obtained by fitting the individual terms of Eq. (3.25) to match predicted and experimental changes in aggregation rates upon mutation.[25] Equation (3.25) was shown to reproduce to a remarkable extent ($r=0.8$) the changes in the aggregation rates observed experimentally for single amino acid substitutions for a series of peptides and proteins, including many associated with disease. It was also shown that other methods can be derived without fitting coefficients, by defining the values of the parameters from general considerations, for instance, according to aromatic contributions, secondary structure propensities, and solvent-accessible areas.[26] The following equation predicts the effect of a mutation on aggregation rate without the use of fitting parameters[26]

$$v_{\mathrm{mut}}/v_{\mathrm{wt}} = \phi_h \phi_\beta \phi_a \phi_c \qquad [3.26]$$

The factor ϕ_h captures most of the apolar and polar interactions. An amino acid is called *p* if its side chain carries a charge or dipole; otherwise, it is called *a*. For mutations that involve the same type of amino acid (a → a) or(p → p)

$$\phi_h = \begin{cases} \mathrm{ASA}^{\mathrm{a}}_{\mathrm{mut}}/\mathrm{ASA}^{\mathrm{a}}_{\mathrm{wt}} & \mathrm{a} \to \mathrm{a} \\ \mathrm{ASA}^{\mathrm{p}}_{\mathrm{wt}}/\mathrm{ASA}^{\mathrm{p}}_{\mathrm{mut}} & \mathrm{p} \to \mathrm{p} \end{cases} \qquad [3.27]$$

where $\mathrm{ASA}^{\mathrm{a}}$ and $\mathrm{ASA}^{\mathrm{p}}$ are the apolar and polar water-accessible surface areas of the amino acid chains.[26]

For mutations that involve different types of amino acids (a → p or p → a)

$$\phi_h = \begin{cases} 1/D_{\mathrm{mut}} & \mathrm{a} \to \mathrm{p} \\ D_{\mathrm{wt}} & \mathrm{p} \to \mathrm{a} \end{cases} \qquad [3.28]$$

where D is the magnitude of the dipole of the amino acid side chains.

The factor ϕ_β is related to the ratio of β-propensity

$$\phi_\beta = \frac{\beta_{\mathrm{mut}}}{\beta_{\mathrm{wt}}} \qquad [3.29]$$

Functions ϕ_a and ϕ_c approximate the effect of aromatic residues A and total charge C

$$\phi_a \phi_c = \exp[\Delta A - \Delta|C|/2] \qquad [3.30]$$

The very high accuracy obtained with these simple mathematical formulas ($r > 0.85$) motivated the development of a series of sequence-based methods.[27,28]

8. PREDICTION OF AGGREGATION RATES

Considering that physicochemical properties of amino acids are important factors for aggregation,[25,26] we and others investigated whether such properties can be used to predict not only the changes in aggregation rates of peptides and proteins upon amino acid substitutions but also the overall aggregation rates starting from the knowledge of their amino acid sequences.[11,29]

In standard *in vitro* experiments, such extrinsic factors include the physicochemical parameters that define the environment of the polypeptides, such as pH, temperature, ionic strength, and protein and denaturant concentrations. Additionally, in order to study the relationship between aggregation and disease, it is important to consider also factors relevant to *in vivo* experiments, including the interactions with cellular components such as molecular chaperones, proteases that generate or process the amyloidogenic precursors, and the effectiveness of quality control mechanisms, as the ubiquitin-proteasome system. All these factors are absent from Eq. (3.25), which therefore is of limited use for the direct prediction of experimentally measured aggregation rates because the intrinsic (i.e., sequence-dependent) aggregation rates will be strongly modulated by extrinsic (i.e., sequence-independent) factors *in vivo*.

The aggregation propensity π_{il} of an l-residue segment starting at position i in a protein sequence can be evaluated as[29]

$$\pi_{il} = \phi_{il}\Phi_{il} \qquad [3.31]$$

The factor Φ_{il} contains exponential functions and is position dependent

$$\Phi_{il} = \exp[A_{il} + B_{il} + C_{il}] \qquad [3.32]$$

where A_{il}, B_{il}, and C_{il} are functions of the aromaticity, β-propensity, and charge. The factor ϕ_{il} depends almost exclusively on the amino acid composition

$$\phi_{il} = \left[\prod_{j=i}^{i+l-1} \left(\frac{S_j^{\mathrm{a}}}{S^{\mathrm{a}}} \theta^{\uparrow\uparrow} + \frac{S_j^{\mathrm{p}}}{S^{\mathrm{p}}} \theta^{\uparrow\downarrow} \right) \frac{S^{\mathrm{t}}}{S_j^{\mathrm{t}}} \frac{\sigma}{\sigma_j} \right] \qquad [3.33]$$

where S_j^{a}, S_j^{p}, S_j^{t}, and σ_j—weighted by their average over the 20 standard amino acids—are the side chain apolar, polar, total water-accessible surface area, and solubility, respectively. The functions $\theta^{\uparrow\uparrow}$ and $\theta^{\uparrow\downarrow}$ include positional effects and reflect the parallel or antiparallel tendency to aggregate if the majority of residues is apolar or polar, respectively.

Considering the high correlation between measured and predicted changes in aggregation rate upon single point mutations,[29] it is possible to utilize π_{il} to predict the absolute rate (Fig. 3.2B)

$$v_{il} = \alpha(c, T)\pi_{il} \tag{3.34}$$

where $\alpha(c, T)$ is introduced to take into account concentration and temperature dependence. Linear relationships between aggregation rates and concentration as well as temperature are assumed in the physiological range.[29]

The aggregation process of peptide and proteins depends strongly on the specific regions of their amino acid sequences whose aggregation propensities are particularly high. The definition of the intrinsic aggregation rate π_{il} enables the aggregation propensity profiles to be calculated in order to identify the aggregation-prone regions.[10,11]

The aggregation propensity profile can be reformulated by introducing the position-dependent score p_i^{agg}. For a given residue i, the p_i^{agg} score is calculated as

$$p_i^{\mathrm{agg}} = \alpha_{\mathrm{h}} p_{\mathrm{h}} + \alpha_{\mathrm{s}} p_{\mathrm{s}} + \alpha_{\mathrm{hyd}} p_{\mathrm{hyd}} \tag{3.35}$$

where p_{h} and p_{s} are the propensities for α-helix and β-sheet formation, respectively, and p_{hyd} is the hydrophobicity.[11] These propensities can be combined in a linear way with coefficients α determined as described below. The p_i^{agg} values are combined to provide a score, A_i^{p}, which describes the intrinsic propensity for aggregation as a function of the complete amino acid sequence.[10,11] At each position i along the sequence, we define the profile A_i^{p} as an average over a window of seven residues

$$A_i^{\mathrm{p}} = \frac{1}{7}\sum_{j=-3}^{3} p_{i+j}^{\mathrm{agg}} + \alpha_{\mathrm{pat}} I_i^{\mathrm{pat}} + \alpha_{\mathrm{gk}} I_i^{\mathrm{gk}} \tag{3.36}$$

where I_i^{pat} is the term that takes into account the presence of specific patterns of alternating hydrophobic and hydrophilic residues[30] and I_i^{gk} is the term that takes into account the gatekeeping effect of individual charges c_i[11]

$$I_i^{\mathrm{gk}} = \sum_{j=-10}^{10} c_{i+j} \tag{3.37}$$

The parameters α were fitted using a Monte Carlo optimization.[10,11] In order to compare the intrinsic propensity profiles, we normalize A_i^{p} by considering the average (μ_{A}) and the standard deviation (σ_{A}) of A_i^{p} at each position i for random sequences. The normalized intrinsic aggregation propensity profile is defined as

$$Z_i^{\mathrm{agg}} = \frac{A_i^{\mathrm{p}} - \mu}{\sigma} \tag{3.38}$$

where we calculated the average μ and the standard deviation σ over random sequences

$$\mu = \frac{1}{(N-8)N_S}\sum_{k=1}^{N_S}\sum_{i=4}^{N-4} A_i^p(S_k), \qquad [3.39]$$
$$\sigma^2 = \frac{1}{(N-8)N_S}\sum_{k=1}^{N_S}\sum_{i=4}^{N-4} \left(A_i^p(S_k)-\mu\right)^2$$

In these formulas, we considered N_S random sequences of length N, and we verified that μ and σ are essentially constant for values of N ranging from 50 to 1000. Random sequences were generated by using the amino acid frequencies of the Uniprot database.

9. PREDICTION OF AGGREGATION-PRONE REGIONS IN NATIVE STATES OF PROTEINS

When a protein is folded, the propensity to form amyloid structures is often inversely related to the stability of its native state.[31] This finding suggests that regions with a high intrinsic propensity for aggregation may be buried inside stable and often highly cooperative structural elements, and therefore unable in such states to form the specific intermolecular interactions that lead to aggregation, although, following mutations that destabilize the native structure, they might acquire this ability.[9] A region of a polypeptide sequence should meet two fundamental conditions in order to promote aggregation: (i) it should have a high intrinsic aggregation propensity and (ii) it should be sufficiently unstructured or unstable to have the opportunity to form intermolecular interactions upon becoming exposed to the solvent through structural fluctuations.[32]

In order to be able to take into consideration the tendency of a given region of a protein sequence to adopt a folded conformation, we introduced the CamP method, which provides a position-dependent score, denoted as ln P_i, predicting the local structural stability at that position.[32] This method enables the high accuracy prediction from the knowledge of amino acid sequence of the regions that are buried in the native state of a protein and of the protection factors for native hydrogen exchange.[32] By combining the predictions of the intrinsic aggregation propensity profiles with those for folding into stable structures, it is possible to account for the influence of the structural context on the aggregation propensities. A new aggregation propensity profile $\widetilde{Z}_i^{\text{agg}}$ can be defined by modulating the intrinsic aggregation propensity profile Z_i^{agg} with the local stability score[32] ln P_i

$$\widetilde{Z}_i^{\text{agg}} = Z_i^{\text{agg}}\left(1 - \frac{\ln P_i}{\ln P_{\max}}\right) \qquad [3.40]$$

where ln $P_{\max}$ is the maximal value that this parameter can reach. These modulations on the Z_i^{agg} profile are made only when $Z_i^{\text{agg}} > 0$ since we consider only the effects on the regions of high intrinsic aggregation propensity, which are those that effectively drive the aggregation process.

From the $\widetilde{Z}_i^{\text{agg}}$ score, it is possible to define an overall aggregation propensity $\widetilde{Z}^{\text{agg}}$ score by summing over all the amino acids of a sequence that have aggregation propensities higher than those of random sequences[11]

$$\widetilde{Z}^{\text{agg}} = \frac{\sum_{i=1}^{N} \widetilde{Z}_i^{\text{agg}} \vartheta\left(\widetilde{Z}_i^{\text{agg}}\right)}{\sum_{i=1}^{N} \vartheta\left(\widetilde{Z}_i^{\text{agg}}\right)} \qquad [3.41]$$

10. LIFE ON THE EDGE—THE ROLE OF PROTEIN CONCENTRATION IN PROMOTING AGGREGATION

Investigating the physicochemical determinants of protein aggregation, we found a remarkable anticorrelation between the expression levels of human genes *in vivo* and the aggregation rates of proteins measured *in vitro*[33] (Fig. 3.3A). A simple principle can be recognized behind our finding that an evolutionary pressure acts to decrease the risk of aggregation in a way that proteins with higher aggregation propensity are generally less expressed. We observe that a "functional amyloid" like *Pmel17*[34] does not follow the trend as no evolutionary pressure is acting on a protein whose aggregation is beneficial for the organism. Thus, aggregation propensities are precisely tuned by evolutionary selection to levels that enable them to be functional at the concentrations required for optimal performance (Fig. 3.3A).

It is possible to speculate on the mathematical representation of the relationship between expression levels and aggregation rates, by defining the quantity

$$p = mv \qquad [3.42]$$

In Eq. (3.42), m represents the expression level and v the aggregation rate. As m and v have practically the same range of variability (Fig. 3.3A), we can assume that $p \approx cst$. In analogy with classical mechanics, p can be regarded as

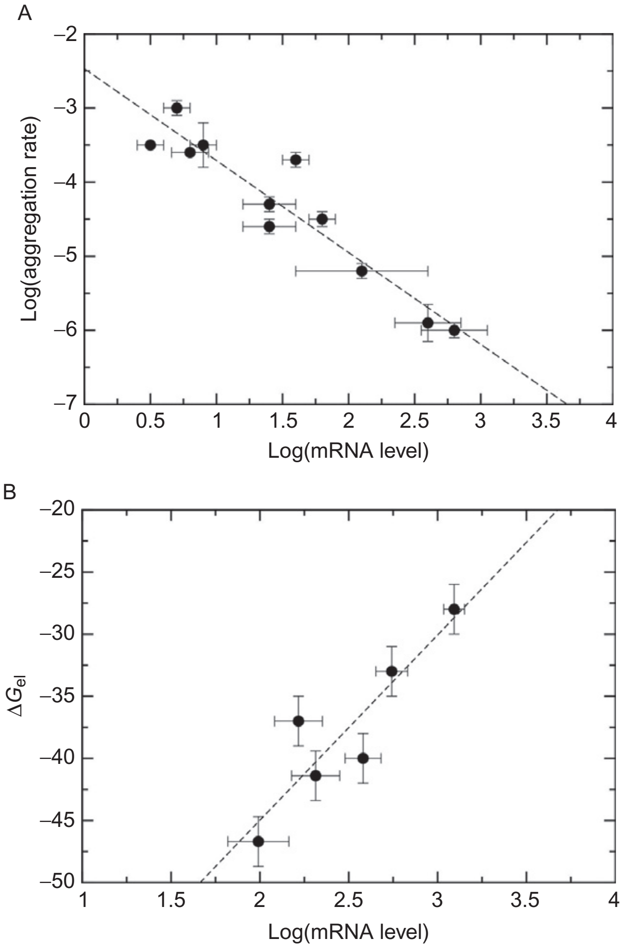

Figure 3.3 Life on the edge. (A) Experimental aggregation rates and mRNA expression levels are strongly anticorrelated. (B) Free energies in the metastable state correlate with mRNA expression levels.

the momentum and $f = dp/dt$ represents the associated force. Since $dp/dt \approx 0$, we can consider the cell in homeostasis. In the presence of perturbations, $dp = v dm + m dv \neq 0$ and additional forces are required to modulate expression levels and aggregation rates. Indeed, the term $v dm$ must be linked to regulatory networks[35,36] and $m dv$ should be associated to molecular chaperones that intervene in order to prevent the formation of nonnative insoluble intermediates when folding into the native state is challenged.[13]

When the concentration M_S of a protein exceeds its critical value (see Eq. 3.6–3.7)[37]

$$M_S^{\max} = \exp[\Delta G_{\text{el}}/KT] \qquad [3.43]$$

the native state is not thermodynamically stable and a protein can in principle lower its overall free energy through amyloid formation, in the same manner in which other types of molecules that exceed their solubility limit have a tendency to form insoluble amorphous or crystalline structures.[37] Do living systems operate under conditions of metastability under normal circumstances? In order to answer this question, we analyzed values from the literature for various critical concentrations.[37] We observe that mRNA expression levels strongly correlate with these critical concentrations (Fig. 3.3B). Our results have profound implications on our understanding of the thermodynamics and kinetics of protein molecules and point to the fact that evolution favors solubility over aggregation.[37]

It should be noted that we assumed a correlation between protein and RNA abundances in our analyses. The correlation between protein concentrations and mRNA expression levels is very well known for bacteria and fungi.[29] However, higher eukaryotes often require substantial posttranscriptional modifications to yield the final amount of protein. To reduce the effect of these modifications in our analysis, we used median scaling and quantile normalization of gene expression levels in different tissues. Accordingly, expression levels were averaged over all the tissues in which a gene was found expressed.[33] This procedure reduces the effect that regulatory processes have on protein expression, because tissue-specific cofactors are averaged out together with environmental conditions.

11. CONCLUSIONS

In this chapter, we described a series of methods for predicting the aggregation of proteins based on their physicochemical properties. The methodology presented is based on the idea that sequences determine protein behavior *in vitro*, in the cases of the folding, misfolding, and aggregation processes, as well as *in vivo*, in the cases of cellular toxicity, solubility, and interactions with chaperones that arise upon protein misfolding.

Our results reveal stringent conditions on the activities of proteins that are dictated by fundamental physicochemical properties. Based on these findings, it is possible to build a theoretical framework to predict which factors contribute most to the aggregation and toxicity of globular proteins, natively unfolded polypeptide chains, and systems that contain both folded and unfolded domains.

A wide number of diseases have been associated to protein misfolding and aggregation. Besides the actual aggregation process, several events that take place both upstream (i.e., mutations, oxidative stress, etc.) and downstream (e.g., promiscuous interactions, chaperones activation) can modify the onset and the severity of such debilitating pathologies, increasing consistently their degree of complexity.[36,38] Quantitative tools are required in order to address such complexity and identify relevance of each factor involved. A theoretical framework like the one proposed in this chapter works in this direction by allowing to describe quantitatively the contribution of the different amino acids to the aggregation process and ultimately to the onset of disease.

Most importantly, the possibility provided by the different theoretical approaches is of significant value in developing rational approaches to avoid aggregation in the biotechnology industry, as well as to understand which are the crucial factors to target in order to prevent this process from happening *in vivo*.

REFERENCES

1. Tandford C, Buckley 3rd CE, De PK, Lively EP. Effect of ethylene glycol on the conformation of gama-globulin and beta-lactoglobulin. *J Biol Chem* 1962;**237**:1168–71.
2. Tanford C. Protein denaturation. C. Theoretical models for the mechanism of denaturation. *Adv Protein Chem* 1970;**24**:1–95.
3. Serpell LC, Sunde M, Benson MD, Tennent GA, Pepys MB, Fraser PE. The protofilament substructure of amyloid fibrils. *J Mol Biol* 2000;**300**:1033–9.
4. Serpell LC. Alzheimer's amyloid fibrils: structure and assembly. *Biochim Biophys Acta* 2000;**1502**:16–30.
5. Fändrich M, Meinhardt J, Grigorieff N. Structural polymorphism of Alzheimer Abeta and other amyloid fibrils. *Prion* 2009;**3**:89–93.
6. Kopito RR. Aggresomes, inclusion bodies and protein aggregation. *Trends Cell Biol* 2000;**10**:524–30.
7. Schlieker C, Bukau B, Mogk A. Prevention and reversion of protein aggregation by molecular chaperones in the E. coli cytosol: implications for their applicability in biotechnology. *J Biotechnol* 2002;**96**:13–21.
8. Kelly JW. The alternative conformations of amyloidogenic proteins and their multi-step assembly pathways. *Curr Opin Struct Biol* 1998;**8**:101–6.
9. Dobson CM. Protein misfolding, evolution and disease. *Trends Biochem Sci* 1999;**24**: 329–32.
10. DuBay KF, Pawar AP, Chiti F, Zurdo J, Dobson CM, Vendruscolo M. Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains. *J Mol Biol* 2004;**341**:1317–26.
11. Tartaglia GG, Pawar AP, Campioni S, Dobson CM, Chiti F, Vendruscolo M. Prediction of aggregation-prone regions in structured proteins. *J Mol Biol* 2008;**380**:425–36.
12. Bukau B, Weissman J, Horwich A. Molecular chaperones and protein quality control. *Cell* 2006;**125**:443–51.
13. Hartl FU, Hayer-Hartl M. Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science* 2002;**295**:1852–8.

14. Schubert U, Antón LC, Gibbs J, Norbury CC, Yewdell JW, Bennink JR. Rapid degradation of a large fraction of newly synthesized proteins by proteasomes. *Nature* 2000;**404**:770–4.
15. Oosawa F, Asakura S. *Thermodynamics of the polymerization of protein.* Waltham, MA, USA: Academic Press Inc.; 1975.
16. Oosawa F, Kasai M. A theory of linear and helical aggregations of macromolecules. *J Mol Biol* 1962;**4**:10–21.
17. Knowles TPJ, Waudby CA, Devlin GL, Cohen SIA, Aguzzi A, Vendruscolo M, et al. An analytical solution to the kinetics of breakable filament assembly. *Science* 2009;**326**: 1533–7.
18. Cohen SIA, Vendruscolo M, Dobson CM, Knowles TPJ. From macroscopic measurements to microscopic mechanisms of protein aggregation. *J Mol Biol* 2012;**421**:160–71.
19. Ferrone FA, Hofrichter J, Sunshine HR, Eaton WA. Kinetic studies on photolysis-induced gelation of sickle cell hemoglobin suggest a new mechanism. *Biophys J* 1980; **32**:361–80.
20. Beaven GH, Gratzer WB, Davies HG. Formation and structure of gels and fibrils from glucagon. *Eur J Biochem* 1969;**11**:37–42.
21. Ruschak AM, Miranker AD. Fiber-dependent amyloid formation as catalysis of an existing reaction pathway. *Proc Natl Acad Sci USA* 2007;**104**:12341–6.
22. Massi F, Straub JE. Energy landscape theory for Alzheimer's amyloid beta-peptide fibril elongation. *Proteins* 2001;**42**:217–29.
23. Kusumoto Y, Lomakin A, Teplow DB, Benedek GB. Temperature dependence of amyloid beta-protein fibrillization. *Proc Natl Acad Sci USA* 1998;**95**:12277–82.
24. Lomakin A, Chung DS, Benedek GB, Kirschner DA, Teplow DB. On the nucleation and growth of amyloid beta-protein fibrils: detection of nuclei and quantitation of rate constants. *Proc Natl Acad Sci USA* 1996;**93**:1125–9.
25. Chiti F, Stefani M, Taddei N, Ramponi G, Dobson CM. Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature* 2003;**424**:805–8.
26. Tartaglia GG, Cavalli A, Pellarin R, Caflisch A. The role of aromaticity, exposed surface, and dipole moment in determining protein aggregation rates. *Protein Sci* 2004;**13**: 1939–41.
27. Sánchez de Groot N, Pallarés I, Avilés FX, Vendrell J, Ventura S. Prediction of "hot spots" of aggregation in disease-linked polypeptides. *BMC Struct Biol* 2005;**5**:18.
28. Fernandez-Escamilla A-M, Rousseau F, Schymkowitz J, Serrano L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol* 2004;**22**:1302–6.
29. Tartaglia GG, Cavalli A, Pellarin R, Caflisch A. Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. *Protein Sci* 2005;**14**: 2723–34.
30. Xiong H, Buckwalter B, Shieh H, Hecht M. Periodicity of polar and nonpolar amino acids is the major determinant of secondary structure in self-assembling oligomeric peptides. *Proc Natl Acad Sci* 1995;**92**:6349–53.
31. Chiti F, Taddei N, Bucciantini M, White P, Ramponi G, Dobson CM. Mutational analysis of the propensity for amyloid formation by a globular protein. *EMBO J* 2000;**19**:1441–9.
32. Tartaglia GG, Cavalli A, Vendruscolo M. Prediction of local structural stabilities of proteins from their amino acid sequences. *Structure* 2007;**15**:139–43.
33. Tartaglia GG, Pechmann S, Dobson CM, Vendruscolo M. Life on the edge: a link between gene expression levels and aggregation rates of human proteins. *Trends Biochem Sci* 2007;**32**:204–6.
34. Fowler DM, Koulov AV, Alory-Jost C, Marks MS, Balch WE, Kelly JW. Functional amyloid formation within mammalian tissue. *PLoS Biol* 2006;**4**:e6.

35. Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* 2004;**431**:308–12.
36. Cirillo D, Agostini F, Klus P, Marchese D, Rodriguez S, Bolognesi B, et al. Neurodegenerative diseases: quantitative predictions of protein-RNA interactions. *RNA* 2013;**19**:129–40.
37. Baldwin AJ, Knowles TPJ, Tartaglia GG, Fitzpatrick AW, Devlin GL, Shammas SL, et al. Metastability of native proteins and the phenomenon of amyloid formation. *J Am Chem Soc* 2011;**133**:14160–3.
38. Johnson R, Noble W, Tartaglia GG, Buckley NJ. Neurodegeneration as an RNA disorder. *Prog Neurobiol* 2012;**99**:293–315.