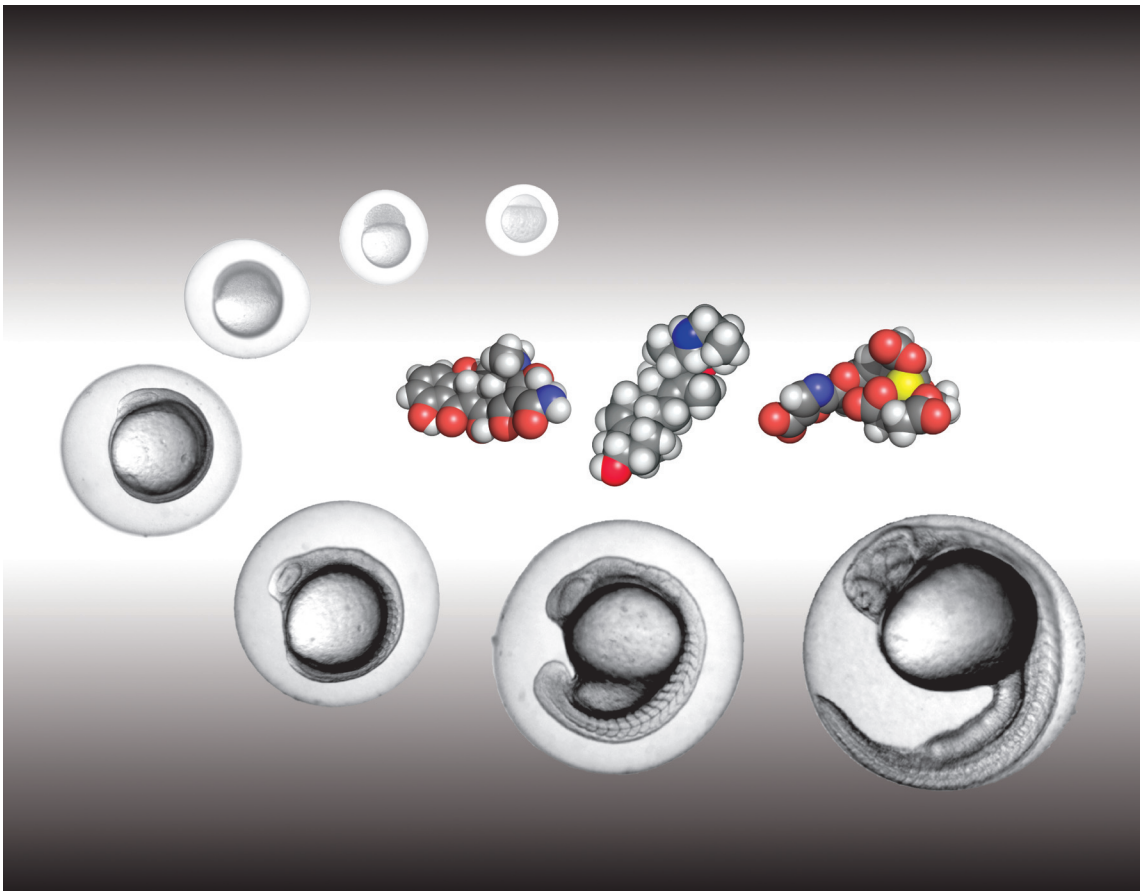
Chem Soc Rev

This article was published as part of the

2008 Chemistry–Biology Interface Issue

Reviewing research at the interface where chemistry meets biology

Please take a look at the full table of contents to access the other papers in this issue



The Zyggregator method for predicting protein aggregation propensities†

Gian Gaetano Tartaglia and Michele Vendruscolo*

Received 15th April 2008
First published as an Advance Article on the web 27th May 2008
DOI: 10.1039/b706784b

Protein aggregation causes many devastating neurological and systemic diseases and represents a major problem in the preparation of recombinant proteins in biotechnology. Major advances in understanding the causes of this phenomenon have been made through the realisation that the analysis of the physico-chemical characteristics of the amino acids can provide accurate predictions about the rates of growth of the misfolded assemblies and the specific regions of the sequences that promote aggregation. More recently it has also been shown that the toxicity *in vivo* of protein aggregates can be predicted by estimating the propensity of polypeptide chains to form protofibrillar assemblies. In this *tutorial review* we describe the development of these predictions made through the Zyggregator method and the applications that have been explored so far.

Introduction

Despite the presence of highly organised cellular processes that regulate the behaviour of proteins *in vivo*,[1] their amino acid sequences play a fundamental role in determining their intrinsic propensities to fold and function,[2] or to misfold and to aggregate.[3,4] Following this observation it has been realised that it is possible to make accurate predictions about whether a protein will aggregate starting from the knowledge of its sequence.[4–15] Thus, considerable progress in understanding and controlling protein aggregation has been made by considering the basic physico-chemical properties of the amino acids.

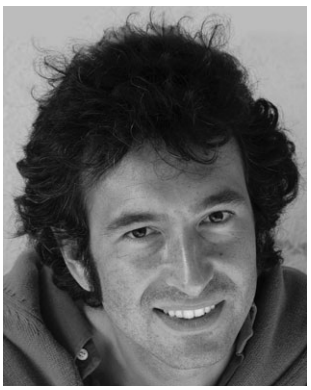
These advances are particularly relevant since protein aggregation into assemblies rich in β-sheet structure has been linked to a series of severe disorders, including Alzheimer's and Parkinson's disease, and type II diabetes.[16–18] Additional interest in this phenomenon comes from the possibility of using highly ordered cross-β protein aggregates known as amyloid fibrils as novel high-performance and versatile nanomaterials,[19] and in reducing the costs caused by protein aggregation into the so called inclusion bodies in the production of proteins for therapeutic use by increasing their solubility.[20]

In this paper we review the development of the Zyggregator algorithm[5,6,21–27] (http://www-vendruscolo.ch.cam.ac.uk/zyggregator.php), a computer program that enables predictions to be made about different phases of the aggregation process and for a variety of experimental conditions.

Changes of aggregation rates upon mutation

The Zyggregator algorithm is based on a seminal study that investigated the role of the physico-chemical properties of amino acids in determining changes in the aggregation rates resulting from individual amino acid substitutions.[4] A significant correlation was found between the changes in the aggregation rates resulting from single mutations and their effect on three physico-chemical properties of the polypeptide chain, hydrophobicity, charge, and the propensity to adopt α-helical or β-sheet structures. These factors were included in an equation to correlate the changes in aggregation rates relative to the wild-type protein for single substitutions in regions of

Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, UK CB2 1EW. E-mail: ggt23@cam.ac.uk.
E-mail: mv245@cam.ac.uk; Fax: +44 1223 763418;
Tel: +44 1223 336366; Tel: +44 1223 763873
† Part of a thematic issue examining the interface of chemistry with biology.



Gian Gaetano Tartaglia studied theoretical physics in Rome. He then received his PhD in biochemistry at the University of Zurich in 2004, and he is currently working as a postdoctoral researcher in Cambridge in the field of protein misfolding and aggregation.

Michele Vendruscolo obtained a PhD in condensed matter physics at SISSA in Trieste (Italy). He is now a lecturer in theoretical chemical biology at the University of Cambridge and an EMBO Young Investigator. His research interests are primarily focused on the investigation of the structures and properties of proteins, and their relationship to misfolding diseases.

the polypeptide chains observed to influence aggregation and for peptides and proteins that were at least partially unfolded

$$\log(k/k') = \alpha_{\mathrm{hydr}}\Delta I^{\mathrm{hydr}} + \alpha_{\mathrm{ss}}\Delta I^{\mathrm{ss}} + \alpha_{\mathrm{ch}}\Delta I^{\mathrm{ch}} \quad (1)$$

In this equation $\log(k/k')$ is the logarithm in base 10 of the ratio of k, the aggregation rate of the wild type and k', the aggregation rate of the mutant, and ΔI^{hydr}, ΔI^{ss} and ΔI^{ch} represent, respectively, the change in hydrophobicity, I^{hydr}, secondary structure propensity, I^{ss}, and charge, I^{ch}, upon mutation. The parameters α were obtained by fitting eqn (1) on a database of mutational variants for which aggregation rates were measured *in vitro*.[4] This formula reproduces to a remarkable extent ($r = 0.85$) the changes in the aggregation rates observed experimentally for single amino acid substitutions for a series of peptides and proteins, including those associated with disease.[4]

Absolute aggregation rates

By exploiting the observation that the physico-chemical properties of amino acids are important factors influencing aggregation, we investigated whether such properties can be used to predict the overall aggregation rates of peptides and proteins starting from the knowledge of their amino acid sequences. On the basis of eqn (1) we considered the following expression for the aggregation rate of a polypeptide chain

$$\log(k) = \alpha_0 + \alpha_{\mathrm{hydr}}I^{\mathrm{hydr}} + \alpha_{\mathrm{ss}}I^{\mathrm{ss}} + \alpha_{\mathrm{ch}}I^{\mathrm{ch}} + \alpha_{\mathrm{pat}}I^{\mathrm{pat}} \quad (2)$$

With respect to eqn (1) we added a term, I^{pat}, to take into account the existence of patterns of alternating hydrophobic and hydrophilic residues, which are known to influence strongly the aggregation process;[28] a factor of +1 is assigned for each pattern of five consecutive alternating hydrophobic and hydrophilic residues in the sequence.[5] As in the case of eqn (1) the α parameters in eqn (2) can be obtained by fitting on a database of proteins for which aggregation rates are measured *in vitro*.[5]

The simple approach of eqn (2), however, does not take into account how diverse factors, extrinsic to the amino acid sequences, influence the rates of aggregation of peptides and proteins. In standard *in vitro* experiments such extrinsic factors include the parameters defining the environment of the polypeptides, such as pH, temperature, ionic strength and protein and denaturant concentrations.[4,18] Additionally, in order to study the relationship between aggregation and disease it is important to consider also factors relevant in *in vivo* experiments such as the interaction 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.[29–31] All these factors are absent from eqn (2), which therefore is of limited use since the fitting of the parameters α in eqn (2) should therefore be carried out with a database of peptides and proteins whose aggregation rates have been measured under identical conditions.

Since in practice it is extremely challenging to construct such a homogenous database of aggregation rates, an approach should be devised that can make use of databases compiled by considering aggregation rates measured under a variety of conditions. A problem to solve is that the physico-chemical propensities considered in eqn (1) and (2) are modified when the experiment are carried out under varying conditions. For example, increasing the concentration of denaturant modifies the hydrophobicity and the secondary structure propensity of the different amino acids. If we consider these modifications to be small we can take them into account by introducing linear corrections

$$\log(k) = \log(k_{\mathrm{int}}) + \log(k_{\mathrm{ext}}) \quad (3)$$

where k_{int} is the "intrinsic" aggregation rate defined by eqn (2) and k_{ext} is an "extrinsic" one. We initially considered the effects of three such factors[5]

$$\log(k_{\mathrm{ext}}) = \alpha_{\mathrm{pH}}E^{\mathrm{pH}} + \alpha_{\mathrm{ionic}}E^{\mathrm{ionic}} + \alpha_{\mathrm{conc}}E^{\mathrm{conc}} \quad (4)$$

where, E^{pH} accounts for the pH of the solution in which aggregation occurs, E^{ionic} defines the ionic strength of the solution, and E^{conc} refers to the polypeptide concentration in the solution.[5]

The parameters α in eqn (3) can be fitted by using a database of aggregation rates determined experimentally under different conditions, at least when such conditions are not too different from the physiological ones.[5] The predictions made through eqn (3) have been tested on a range of peptides and proteins, providing accurate predictions ($r = 0.8$ or better) for aggregation rates spanning over five orders of magnitude,[5] thus showing that it is possible to rationalise the aggregation process *in vitro* on the basis of relatively simple combination of physico-chemical properties of the amino acid sequences and of the environment in which they are found.

We observe that the predictions of the changes of aggregation rates upon mutation made using eqn (1) are rather accurate even if they only consider intrinsic factors, since the ratio $k_{\mathrm{int}}/k_{\mathrm{int}}*$ is, according to eqn (3), equal to $k/k*$ which are the rates actually observed in the experiment.

Aggregation-prone regions

The aggregation process of peptide and proteins depends strongly on the specific regions of their amino acids sequences whose aggregation propensities are particularly high.[6–15] The definition of the intrinsic aggregation rate k_{int} enables aggregation propensity profiles to be calculated in order to identify these aggregation-prone regions.[6]

The aggregation propensity profile is defined by considering the position-dependent P_i^{agg} score.[6,26] For a given residue i, the P_i^{agg} score is calculated as[26]

$$P_i^{\mathrm{agg}} = \frac{1}{7}\sum_{j=-3}^{3} p_{i+j}^{\mathrm{agg}} + \alpha_{\mathrm{pat}}I_i^{\mathrm{pat}} + \alpha_{\mathrm{gk}}I_i^{\mathrm{gk}} \quad (5)$$

where we considered the aggregation rate of a seven-residue segment of the protein centered at position i. In eqn (5) the intrinsic aggregation propensity, p_i^{agg}, of an individual amino acid is defined as[26]

$$p_i^{\mathrm{agg}} = \alpha_h p_i^h + \alpha_s p_i^s + \alpha_{\mathrm{hyd}}p_i^{\mathrm{hyd}} + \alpha_c p_i^c \quad (6)$$

where p_i^h, p_i^s, p_i^{hyd}, p_i^c are the amino acid scales for α-helix and β-sheet formation, hydrophobicity and charge.[26] The

remaining two terms in eqn (5), I_i^{pat} and I_i^{gk}, are included, respectively, to account for the presence of hydrophobic patterns and of gatekeeper residues.[32,33] The term I_i^{pat} is 1 if residue i is included in a hydrophobic pattern and 0 otherwise, while the term I_i^{gk} is defined as[26]

$$I_i^{gk} = \sum_{j=-10}^{10} c_{i+j} \quad (7)$$

where the sum over the charges c_i of individual amino acids is made over a sliding window of 21 residues; shorter windows are considered at the N- and C-termini. The term I_i^{gk} is introduced to take into account the fact that when a hydrophobic pattern is flanked by charged residues its contribution to the aggregation propensity is much reduced by electrostatic repulsions.

The P_i^{agg} score is normalised in order to facilitate the comparison between amino acid sequences of different lengths[14]

$$Z_i^{agg} = \frac{P_i^{agg} - \mu^{agg}}{\sigma^{agg}} \quad (8)$$

where the average μ^{agg}

$$\mu^{agg} = \frac{1}{(N-6)\cdot N_S} \sum_{k=1}^{N_S} \sum_{i=4}^{N-3} Z_i^{agg}(S_k) \quad (9)$$

and standard deviation σ^{agg}

$$\sigma^{agg} = \sqrt{\frac{1}{(N-6)\cdot N_S} \sum_{k=1}^{N_S} \sum_{i=4}^{N-3} (Z_i^{agg}(S_k) - \mu^{agg})^2} \quad (10)$$

are calculated over N_S random sequences (with $N_S = 1000$) of length N generated by using the amino acid frequencies of the SWISS-PROT database.[26] With this normalisation, the Z_i^{agg} score is 0 if the aggregation propensity at position i along the sequence is equal to that of a random sequence and 1 if it is one standard deviation more aggregation-prone.

From the Z_i^{agg} score it is possible to define an overall aggregation propensity by summing over all the amino acids of a sequence that have aggregation propensities higher than those of random sequences[26]

$$Z^{agg} = \frac{\sum_{i=1}^{N} Z_i^{agg} \vartheta(Z_i^{agg})}{\sum_{i=1}^{N} \vartheta(Z_i^{agg})} \quad (11)$$

where the function $\vartheta(Z_i^{agg})$ is 1 for $\vartheta(Z_i^{agg}) \geq 0$ and 0 for $Z_i^{agg} < 0$.

The Z_i^{agg} profiles enable a variety of experimental observations about the amyloidogenic potential of different regions of a polypeptide sequence to be rationalised, at least in the cases in which peptides and proteins aggregate from disordered states under physiological conditions. We discuss here the cases of Aβ and α-synuclein.

$A\beta_{1-42}$

The amyloid β-peptide (Aβ) is the main constituent of the extracellular deposits characteristic of Alzheimer's disease (AD).[34] This peptide is found in the human brain predominately in two forms, of 40- and 42-amino acids in length ($A\beta_{1-40}$ and $A\beta_{1-42}$, respectively).

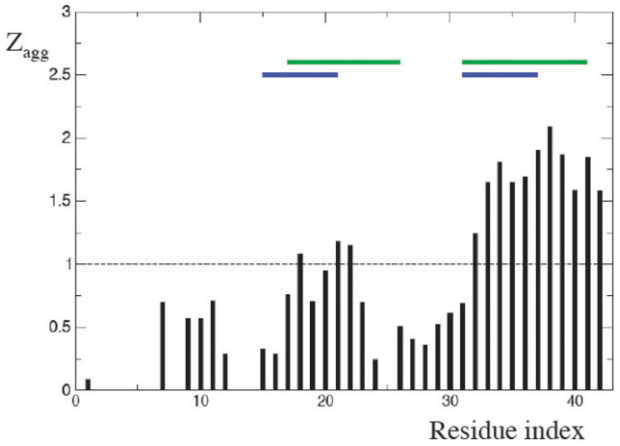


Fig. 1 Aggregation propensity profile of $A\beta_{1-42}$. The vertical bars indicate the intrinsic aggregation propensity profile, Z_i^{agg}. The green and blue horizontal bars indicate regions of the sequence found to form the core of the fibrils as determined with NMR measurement[35,36] and site-directed spin labelling coupled to EPR,[37] respectively.

The intrinsic aggregation propensity profile, Z_i^{agg}, of $A\beta_{1-42}$ reveals two regions of high aggregation propensity (those above the $Z_i^{agg} = 1$ threshold, dashed line in Fig. 1): the central (residues 18–22) and the C-terminal (residues 32–42) regions. Both these regions play an important structural role in the current models of the structures of the $A\beta_{1-40}$ and $A\beta_{1-42}$ peptides in their amyloid forms.

α-Synuclein

Human α-synuclein is known to self-assemble into intracellular inclusions in dopaminergic neurons of patients suffering from Parkinson's disease.[38] Using an array of experimental techniques, including limited proteolysis,[39,40] hydrogen–deuterium exchange[41] and site-directed spin labelling/EPR[42,43] it was found that the central region (approximately residues 30–95) of this normally natively unfolded protein forms the core of the fibrils. The aggregation propensity profile Z_i^{agg} identifies four peaks located within this central region of the sequence (Fig. 2). These four peaks appear to correspond to the regions found to form the β-core of the fibrils using solid-state NMR measurements.

Aggregation-prone regions in the presence of denaturants

In order to obtain an expression for the aggregation propensity profiles that is also valid under strongly non-physiological conditions, we should consider scales of physico-chemical factors determined under such conditions. The approach that we followed in eqn (3) was based on the assumption that the complex dependencies of intrinsic and extrinsic factors can be captured by linear expressions. For weak perturbations[5,6] this approximation is rather accurate, but under harsher conditions we do not expect this to be the case. For example, the

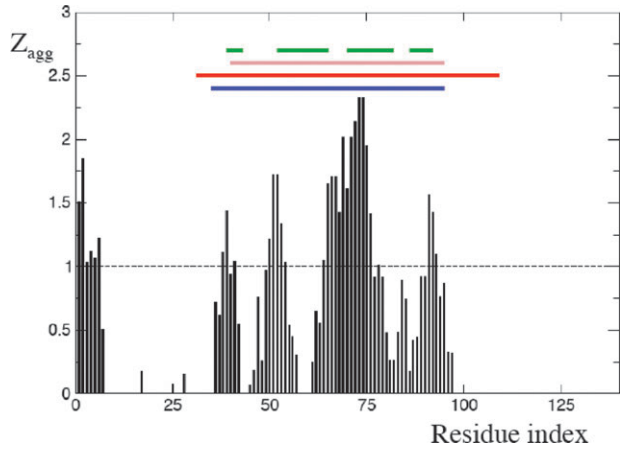


Fig. 2 Aggregation propensity profile of human α-synuclein. The vertical bars indicate the intrinsic aggregation propensity profile, Z_i^{agg}. The blue, red and pink horizontal bars indicate the region of the sequence that appears to be structured in the fibrils from site-directed spin labelling coupled to EPR,[42,43] hydrogen–deuterium exchange[41] and limited proteolysis respectively.[39,40] The regions adopting a β-strand conformation within such a region, as revealed by solid-state NMR measurements, are indicated by green horizontal bars; due to experimental uncertainties the boundaries of such strands are approximate.

addition of TFE changes rather dramatically the secondary structure propensities of individual amino acids. These changes are unlikely to be accurately accounted for by adding up extrinsic contributions to the corresponding intrinsic scales used in eqn (1).

The strategy that we have used in the Zyggregator algorithm to carry out predictions of aggregation propensities in the presence of TFE is to refit the parameters of eqn (2) using a database of *in vitro* aggregation rates for a set of polypeptide chains whose aggregation was monitored in the presence of TFE. We thus obtained a Z_i^{TFE} score analogous to the Z_i^{agg} score defined in eqn (8) but applicable to cases in which aggregation takes place in the presence of TFE.[27]

Aggregation-prone regions in globular states

When a protein is folded, the propensity to form amyloid structures is often inversely related to the stability of its native state.[44] This finding suggests that regions with a high intrinsic propensity for aggregation are 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. A region of a polypeptide sequence should meet two conditions in order to promote aggregation: (1) it should have a high intrinsic aggregation propensity and (2) it should be sufficiently unstructured or unstable to have a significant propensity to form intermolecular interactions.

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 use the CamP method, which provides a position-dependent score, denoted as $\ln P_i$, that characterises the local stability at that position.[45] 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.[45]

By combining the predictions of the intrinsic aggregation propensity profiles with those for folding into stable structures, we account for the influence of the structural context on the aggregation propensities. We thus define[26] a new aggregation propensity profile $\tilde{Z}_i^{agg}$, by modulating the intrinsic aggregation propensity profile, Z_i^{agg} with the local stability score, $\ln P_i$

$$\tilde{Z}_i^{agg} = Z_i^{agg}\left(1 - \frac{\ln P_i}{15}\right) \qquad (12)$$

These modulations on the Z_i^{agg} profile are made only when $Z_i^{agg} > 0$ 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 $\tilde{Z}_i^{agg}$ score it is possible to define an overall aggregation propensity by summing over all the amino acids of a sequence that have aggregation propensities higher than those of random sequences[26]

$$\tilde{Z}^{agg} = \frac{\sum_{i=1}^{N} \tilde{Z}_i^{agg}\vartheta(\tilde{Z}_i^{agg})}{\sum_{i=1}^{N} \vartheta(\tilde{Z}_i^{agg})} \qquad (13)$$

We illustrate this approach in the case of the human prion protein (hPrP), which is involved in sporadic, inherited or infectious forms of Creutzfeldt-Jakob disease (CJD), Gerstmann-Straussler-Sheinker disease (GSS) and fatal familial insomnia (FFI).[46] The key event in the pathogenesis of these human diseases is the conversion of the normal α-helical protease-sensitive cellular form of the prion protein ($hPrP^C$) into a β-rich form ($hPrP^{Sc}$) that possesses distinct features such as protease resistance, insolubility and toxicity.[47] Furthermore, $hPrP^{Sc}$ itself appears to mediate the transmission of TSEs by promoting the conversion of $hPrP^C$ into its modified and pathogenic aggregated state.

While the mechanism of conversion of $hPrP^C$ to $hPrP^{Sc}$ is not known in detail, specific regions of the $hPrP^C$ sequence appear to be particularly important in modulating the interaction with $hPrP^{Sc}$ and promoting the process of amyloid formation.[47] In Fig. 3 we show the intrinsic aggregation propensity profile Z_i^{agg} for the sequence of hPrP(23–231). We then took into account the effects of the intrinsic propensities of the various residues to be structured, and hence protected from aggregation resulting in the $\tilde{Z}_i^{agg}$ profile. The similarity of the Z_i^{agg} and the $\tilde{Z}_i^{agg}$ profiles for residues 23–125 is in agreement with the experimental observation that this region is not structured.[48] When considering both intrinsic sequence-based propensities and specific structural factors, the region spanning residues 120–126 corresponds to the highest peak in the entire sequence and the only one to have $\tilde{Z}_i^{agg} > 1$, suggesting that this region is the most aggregation-prone region in the $hPrP^C$ form. This prediction correlates well with experimental data on the *in vitro* aggregation behaviour of

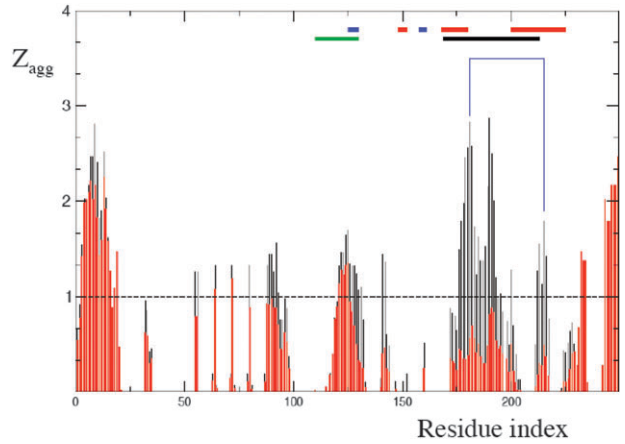


Fig. 3 Aggregation propensity profiles of the human prion protein. The black vertical bars indicate the intrinsic aggregation propensity profile, Z_i^{agg}; the red vertical bars indicate the aggregation propensity profile, $\tilde{Z}_i^{agg}$, calculated by taking into account the structural protection provided by the globular structure of $hPrP^C$ form of the protein, as predicted by the $\ln P_i$ score.[26,45] For reference, the secondary structure elements present in $hPrP^C$ are indicated as blue bars (β-strands) and red bars (α-helices),[48] and the position of the disulfide bond C179–C214 is indicated by a blue line. An experimentally-determined aggregation-prone fragment[47] (residues 118–128) is indicated by a green bar, and it is shown to overlap substantially with the major region predicted by our method to have a significant aggregation propensity ($\tilde{Z}_i^{agg} > 1$) in the $hPrP^C$ form. The region corresponding to the structural core of the amyloid fibril as determined by hydrogen–deuterium exchange[49] (residues 169–213) is indicated by a black bar, and corresponds to the region of high intrinsic aggregation propensity ($Z_i^{agg} > 1$) formed by residues 175–193.

hPrP fragments. Peptides $hPrP_{106–114}$, $hPrP_{106–126}$, $hPrP_{113–126}$ and $hPrP_{127–147}$ of recombinant hPrP all have high propensities to form amyloid fibrils.[47]

Crucially, the aggregation propensity of the region 175–193, which includes α-helix II in the $hPrP^C$ form, is predicted to be very high. Indeed, the intrinsic aggregation propensity profile Z_i^{agg} (Fig. 3) identifies this region as the most amyloidogenic one. However, using the CamP method we also predicted, in agreement with experimental data, that this region is highly structured in the $hPrP^C$ form.[48] Therefore when the aggregation propensity profile $\tilde{Z}_i^{agg}$ is considered, the region of residues 175–193 results to be less aggregation-prone in the $hPrP^C$ form than the region of residues 118–128. In addition, the presence of the disulfide bond C179–C214 appears to play an important role in stabilizing the region 175–193 in the $hPrP^C$ form by inhibiting the formation of intermolecular interactions from this state. However, recent hydrogen/deuterium exchange experiments[49] have indicated that the region corresponding to the structural core of the amyloid fibril corresponds to residues 169–213 (black bar in Fig. 3), a result in close agreement with our predictions using the intrinsic aggregation propensity profile Z_i^{agg}. Therefore the comparison of the Z_i^{agg} and $\tilde{Z}_i^{agg}$ profiles suggests that the region of residues 175–193 is involved in the stabilization of the $hPrP^{Sc}$ forms after the $hPrP^C$ form has been destabilized.

Toxicity of protein assemblies

A question of central importance is whether the possibility of predicting aggregation rates based on the physico-chemical properties of the amino acids is relevant to understand the causes of the toxicity of the aggregates. It is thus crucial to understand the relationship between the toxicity of misfolded assemblies measured *in vivo*, the aggregation rates measured *in vitro*, and the aggregation propensities estimated by computational methods.

We investigated this relationship by carrying out experiments on a transgenic *Drosophila* model of Alzheimer's diseases.[21] By designing a series of mutational variants of the Aβ peptide we established a link between the physico-chemical properties of the sequences of the peptides and the conditions of the flies expressing them in the central nervous system.[21] Since increasing evidences suggests that the most toxic protein aggregates are β-rich oligomeric assemblies known as protofibrillar species,[50–53] we defined a position-dependent toxicity score, Z_i^{tox}, that accounts for the propensity to form protofibrillar assemblies[21]

$$Z_i^{tox} = \frac{P_i^{tox} - \mu^{tox}}{\sigma^{tox}} \qquad (14)$$

In this equation, the terms contributing to Z_i^{tox} are the same as in eqn (8), but with the difference that the parameters are fitted on a database of polypeptide chains whose aggregation resulted in protofibrillar species, rather than amyloid fibrils.[21] The Z_i^{tox} is shown in the case of $A\beta_{1–42}$ in Fig. 4.

In order to compare the predictions with the experimental results we defined an over toxicity score as[21]

$$Z^{tox} = \sum_{i=1}^{N} Z_i^{tox} \qquad (15)$$

The correlation between the Z^{tox} score and the toxicity of the of $A\beta_{1–42}$ mutants was found to be very high ($r = 0.83$) and better than the correlation obtained with the aggregation propensity score Z^{agg} ($r = 0.75$), thus supporting the

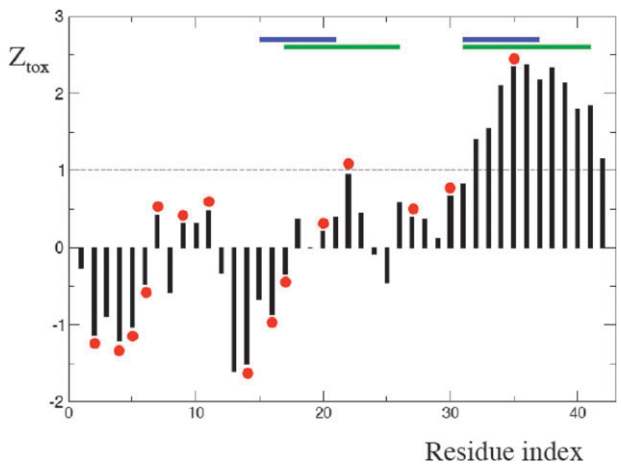


Fig. 4 Toxicity profile of $A\beta_{1–42}$.The vertical bars indicate the toxicity profile, Z_i^{tox}. The positions of the mutants whose toxicity has been studied *in vivo*[21] are indicated by red circles. See Fig. 1 for the definition of the blue and green bars.

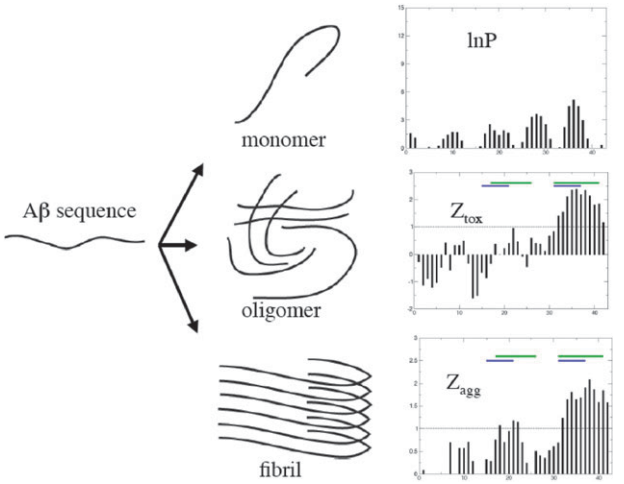


Fig. 5 Strategy of the Zyggregator predictions. We predict different propensities from the amino acid sequence of a peptide or protein: local stability in the monomeric state ($\ln P_i$), formation of β-rich oligomers (Z_i^{tox}), or formation of fibrillar aggregates (Z_i^{agg}). The specific predictions are shown here for $A\beta_{1-42}$.

observation that protofibrillar species are more toxic than fibrillar ones.[21]

The strategy of the Zyggregator predictions

The predictions made with Zyggregator are based on the possibility to estimate whether a peptide or protein will fold or aggregate into fibrillar or protofibrillar structures (Fig. 5) on the basis of combinations of physico-chemical properties of its amino acids (Table 1). For each of these possible outcomes, a different propensity is calculated by constructing a different predictor through a fitting procedure that exploits the experimental knowledge of the rates of the corresponding process, in this case either folding or aggregation into oligomers or fibrils (Fig. 5).

Since the physico-chemical properties of the amino acids change with the environment in which the folding or aggregation processes take place, the coefficients are fitted on a database of experimental rates collected for relatively homogeneous conditions, and the predictions are made only for in the vicinity of such conditions.

Starting from the amino acid sequence of a peptide or protein, the major parameters that determine the propensity for aggregation or for being locally stable in the folded state are calculated from the physico-chemical properties of the amino acids.

Table 1 Parameters included in the Zyggregator predictions

α-Helix propensity	I_i^h
β-Sheet propensity	Ih_i^s
Hydrophobicity	I_i^{hyd}
Charge	I_i^c
Hydrophobic patterns	I_i^{pat}
Gatekeepers	I_i^{gk}
Local stability	$\ln P_i$

Relationship with other methods of predicting protein aggregation propensities

Since the initial realisation that protein aggregation propensities of peptides and proteins can be predicted from the physico-chemical properties of their amino acid sequences,[4] several sequence-based methods have been proposed to achieve this goal.[5–11] These methods differ in the specific way in which the properties of amino acids are translated into phenomenological terms describing the different contributions to the overall propensity for aggregation. For example, in addition to the terms described in eqn (1), the TANGO method considers explicitly the enthalpic and entropic costs associated to the conformational transition between folded and aggregated structures,[7] and the method by Tartaglia *et al.* includes a term to describe the π-stacking contributions to the stability of the aggregates.[8]

More recently, it has also been realised that the aggregation propensities of polypeptide chains can be predicted by following two conceptually distinct strategies. In the first, amino acid sequences are threaded on known cross-β structures, in order to assess its compatibilty with this type of conformation.[12] In the second, the propensities of polypeptide chain to self-assemble into ordered cross-β aggregates are estimated by constructing a knowledge-based residue-residue interaction potential using a database of native structures.[15]

These results indicate that there are currently at least three alternative, almost equivalent strategies for predicting the aggregation propensities of peptides and proteins. Such propensities can be estimated either from to the physico-chemical properties of the amino acid sequences,[4] or according to their compability with the cross-β motif typical of ordered fibrillar assemblies,[12–14] or by considering their tendency of forming β structures in native states.[15] These results strongly support the view that folding and aggregation are two closely related processes that depend primarily on the fundamental physico-chemical properties of polypeptide chains.

Conclusions

We have described the Zyggregator approach for predicting the aggregation propensities of polypeptide chains based on their amino acid sequences. The methodology that we have presented is based on the idea that the sequence of a protein determines its behaviour in the case of folding, misfolding and aggregation.

The possibility provided by methods such as the one that we have presented to predict the regions most important to cause aggregation and toxicity for natively unfolded polypeptide chains, for globular proteins and for systems that contain both folded and unfolded domains should be of significant value in developing rational approaches to the avoidance of aggregation in biotechnology and to the treatment of protein deposition diseases.

References

1. L. H. Hartwell, J. J. Hopfield, S. Leilber and A. W. Murray, *Nature*, 1999, **402**, C47–C52.
2. C. B. Anfinsen, *Science*, 1973, **181**, 223–230.

3. C. M. Dobson, *Trends Biochem. Sci.*, 1999, **24**, 329–332.
4. F. Chiti, M. Stefani, N. Taddei, G. Ramponi and C. M. Dobson, *Nature*, 2003, **424**, 805–808.
5. K. F. Dubay, A. P. Pawar, F. Chiti, J. Zurdo, C. M. Dobson and M. Vendruscolo, *J. Mol. Biol.*, 2004, **341**, 1317–1326.
6. A. P. Pawar, K. F. DuBay, J. Zurdo, F. Chiti, M. Vendruscolo and C. M. Dobson, *J. Mol. Biol.*, 2005, **350**, 379–392.
7. A. M. Fernandez-Escamilla, F. Rousseau, J. Schymkowitz and L. Serrano, *Nat. Biotechnol.*, 2004, **22**, 1302–1306.
8. G. G. Tartaglia, A. Cavalli, R. Pellarin and A. Caflisch, *Protein Sci.*, 2005, **13**, 1939–1941.
9. O. V. Galzitskaya, S. O. Garbuzynskiy and M. Y. Lobanov, *PLoS Comp. Biol.*, 2007, **2**, 1639–1648.
10. O. Conchillo-Sole, N. S. de Groot, F. X. Aviles, J. Vendrell, X. Daura and S. Ventura, *BMC Bioinf.*, 2007, **8**, 65.
11. S. Zibaee, O. S. Makin, M. Goedert and L. C. Serpell, *Protein Sci.*, 2007, **16**, 906–918.
12. M. J. Thompson, S. A. Sievers, J. Karanicolas, M. I. Ivanova, D. Baker and D. Eisenberg, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 4074–4078.
13. Z. Q. Zhang, H. Chen and L. H. Lai, *Bioinformatics*, 2007, **17**, 2218–2225.
14. M. Cecchini, R. Curcio, M. Pappalardo, R. Melki and A. Caflisch, *J. Mol. Biol.*, 2006, **357**, 1306–1321.
15. A. Trovato, F. Chiti, A. Maritan and F. Seno, *PLoS Comp. Biol.*, 2007, **2**, 1608–1618.
16. D. J. Selkoe, *Nature*, 2003, **426**, 900–904.
17. C. M. Dobson, *Nature*, 2003, **426**, 884–890.
18. F. Chiti and C. M. Dobson, *Annu. Rev. Biochem.*, 2006, **75**, 333–366.
19. T. P. Knowles, A. F. Fitzpatrick, S. Meehan, H. Mott, M. Vendruscolo, C. M. Dobson and M. E. Welland, *Science*, 2007, **318**, 1900–1903.
20. S. Ventura and A. Villaverde, *Trends Biotechnol.*, 2006, **24**, 179–185.
21. L. M. Luheshi, G. G. Tartaglia, A. C. Brorsson, A. P. Pawar, I. E. Watson, F. Chiti, M. Vendruscolo, D. A. Lomas, C. M. Dobson and D. C. Crowther, *PLoS Biol.*, 2007, **5**, 2495–2500.
22. J. Meinhardt, G. G. Tartaglia, A. P. Pawar, T. Christopeit, P. Hortschansky, V. Schroeckh, C. M. Dobson, M. Vendruscolo and M. Fandrich, *Protein Sci.*, 2007, **16**, 1214–1222.
23. E. Monsellier, M. Ramazzotti, P. Polverino de Laureto, G. G. Tartaglia, N. Taddei, A. Fontana, M. Vendruscolo and F. Chiti, *Biophys. J.*, 2007, **93**, 4382–4391.
24. S. Meehan, A. J. Baldwin, T. P. J. Knowles, J. F. Smith, A. M. Squires, P. Clements, T. M. Treweek, H. Ecroyd, G. G. Tartaglia, M. Vendruscolo, C. E. MacPhee, C. M. Dobson and J. A. Carver, *J. Mol. Biol.*, 2007, **372**, 470–484.
25. R. C. Rivers, J. R. Kumita, M. M. Dedmon, A. Pawar, G. G. Tartaglia, M. Vendruscolo, C. M. Dobson and J. Christodoulou, *Protein Sci.*, 2008, **17**, 887–898.
26. G. G. Tartaglia, A. P. Pawar, S. Campioni, C. M. Dobson, F. Chiti and M. Vendruscolo, *J. Mol. Biol.*, 2008, in press.
27. M. Calamai, G. G. Tartaglia, M. Vendruscolo, F. Chiti and C. M. Dobson, unpublished work.
28. B. M. Broome and M. H. Hecht, *J. Mol. Biol.*, 2000, **296**, 961–968.
29. N. F. Bence, R. M. Sampat and R. R. Kopito, *Science*, 2001, **292**, 1552–1555.
30. G. K. Tofaris, A. Razzaq, B. Ghetti, K. S. Lilley and M. G. Spillantini, *J. Biol. Chem.*, 2003, **298**, 44405–44411.
31. K. S. McNaught and C. W. Olanow, *Ann. Neurol.*, 2006, **60**, 243–247.
32. D. E. Otzen, O. Kristensen and M. Oliveberg, *Proc. Natl. Acad. Sci. U. S. A.*, 2000, **97**, 9907–9912.
33. F. Rousseau, L. Serrano and J. W. H. Schymkowitz, *J. Mol. Biol.*, 2004, **355**, 1037–1047.
34. D. J. Selkoe, *Science*, 2002, **298**, 789–791.
35. A. T. Petkova, Y. Ishii, J. J. Balbach, O. N. Antzutkin, R. D. Leapman, F. Delaglio and R. Tycko, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 16742–16747.
36. T. Luhrs, C. Ritter, M. Adrian, D. Riek-Loher, B. Bohrmann, H. Doeli, D. Schubert and R. Riek, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 17342–17347.
37. M. Torok, S. Milton, R. Kayed, P. Wu, T. McIntire, C. G. Glabe and R. Langen, *J. Biol. Chem.*, 2002, **277**, 40810–40815.
38. M. H. Polymeropoulos, C. Lavedan, E. Leroy, S. E. Ide, A. Dehejia, A. Dutra, B. Pike, H. Root, J. Rubenstein, R. Boyer, E. S. Stenroos, S. Chandrasekharappa, A. Athanassiadou, T. Papapetropoulos, W. G. Johnson, A. M. Lazzarini, R. C. Duvoisin, G. Di Iorio, L. I. Golbe and R. L. Nussbaum, *Science*, 1997, **276**, 2045–2047.
39. H. Miake, H. Mizusawa, T. Iwatsubo and M. Hasegawa, *J. Biol. Chem.*, 2002, **277**, 19213–19219.
40. Z. Qin, D. Hu, S. Han, D. P. Hong and A. L. Fink, *Biochemistry*, 2007, **46**, 13322–13330.
41. C. Del Mar, E. A. Greenbaum, L. Mayne, S. W. Englander and V. L. Woods, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 15477–15482.
42. A. Der-Sarkissian, C. C. Jao, J. Chen and R. Langen, *J. Biol. Chem.*, 2003, **278**, 37530–37535.
43. M. Chen, M. Margittai, J. Chen and R. Langen, *J. Biol. Chem.*, 2007, **282**, 24970–24979.
44. F. Chiti, N. Taddei, M. Bucciantini, P. White, G. Ramponi and C. M. Dobson, *EMBO J.*, 2000, **19**, 1441–1449.
45. G. G. Tartaglia, A. Cavalli and M. Vendruscolo, *Structure*, 2007, **15**, 139–143.
46. S. B. Prusiner, *Science*, 1991, **252**, 1515–1522.
47. G. Forloni, N. Angeretti, R. Chiesa, E. Monzani, M. Salmona, O. Bugiani and F. Tagliavini, *Nature*, 1993, **362**, 543–546.
48. R. Zahn, A. Z. Liu, T. Luhrs, R. Riek, C. von Schroetter, F. L. Garcia, M. Billeter, L. Calzolai, G. Wider and K. Wuthrich, *Proc. Natl. Acad. Sci. U. S. A.*, 2000, **97**, 145–150.
49. X. J. Lu, P. L. Wintrode and W. K. Surewicz, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 1510–1515.
50. M. P. Lambert, A. K. Barlow, B. A. Chromy, C. Edwards, R. Freed, M. Liosatos, T. E. Morgan, I. Rozovsky, B. Trommer, K. L. Viola, P. Wals, C. Zhang, C. E. Finch, G. A. Krafft and W. L. Klein, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 6448–6453.
51. D. M. Walsh, I. Klyubin, J. V. Fadeeva, W. K. Cullen, R. Anwyl, M. S. Wolfe, M. J. Rowan and D. J. Selkoe, *Nature*, 2002, **416**, 535–539.
52. J. P. Cleary, D. M. Walsh, J. J. Hofmeister, G. M. Shankar, M. A. Kuskowski, D. J. Selkoe and K. H. Ashe, *Nat. Neurosci.*, 2005, **8**, 79–84.
53. C. Haass and D. J. Selkoe, *Nat. Rev. Mol. Cell Biol.*, 2007, **8**, 101–112.