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,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/zygregator.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
the polypeptide chains observed to influence aggregation and for peptides and proteins that were at least partially unfolded
\[
\log(k/k') = \log(k) = \log(k_{int}) + \log(k_{ext})
\]
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
\[
\log(k_{ext}) = \log(k_{int}) + \log(k_{ext})
\]
where \(E_{ext}\) accounts for the pH of the solution in which aggregation occurs, \(E_{conc}\) defines the ionic strength of the solution, and \(E_{conc}\) refers to the polypeptide concentration in the solution.5

Aggregation-prone regions
The aggregation process of peptide and proteins depends strongly on the specific regions of their amino acid 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 \(P_{agg}\) score.6 For a given residue \(i\), the \(P_{agg}\) score is calculated as
\[
P_{agg} = \frac{1}{3} \left( \text{agg score} + \text{agg score} + \text{agg score} \right)
\]
remaining two terms in eqn (5), $I^\text{pat}$ and $I^\text{sk}$, are included, respectively, to account for the presence of hydrophobic patterns and of gatekeeper residues. The term $I^\text{pat}$ is 1 if residue $i$ is included in a hydrophobic pattern and 0 otherwise, while the term $I^\text{sk}$ is defined as:

$$I^\text{sk} = \sum_{j=1}^{10} c_{i+j}$$  \hspace{1cm} (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 $I^\text{sk}$ 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 $P_i^{\text{agg}}$ score is normalised in order to facilitate the comparison between amino acid sequences of different lengths

$$Z_i^{\text{agg}} = \frac{P_i^{\text{agg}} - \mu^{\text{agg}}}{\sigma^{\text{agg}}}$$  \hspace{1cm} (8)

where the average $\mu^{\text{agg}}$ and standard deviation $\sigma^{\text{agg}}$

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

$$\sigma^{\text{agg}} = \sqrt{\frac{1}{(N-6) \cdot N_S} \sum_{k=1}^{N_S} \sum_{i=4}^{N-3} (Z_i^{\text{agg}}(S_k) - \mu^{\text{agg}})^2}$$  \hspace{1cm} (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. With this normalisation, the $Z_i^{\text{agg}}$ 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 $Z_i^{\text{agg}}$ 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

$$Z^{\text{agg}} = \frac{\sum_{i=1}^{N} Z_i^{\text{agg}} \theta(Z_i^{\text{agg}})}{\sum_{i=1}^{N} \theta(Z_i^{\text{agg}})}$$  \hspace{1cm} (11)

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

The $Z_i^{\text{agg}}$ 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.

Aβ

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

The intrinsic aggregation propensity profile, $Z_i^{\text{agg}}$, of Aβ$_{1-42}$ reveals two regions of high aggregation propensity (those above the $Z_i^{\text{agg}} = 1$ 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 Aβ$_{1-40}$ and Aβ$_{1-42}$ 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. Using an array of experimental techniques, including limited proteolysis, hydrogen–deuterium exchange and site-directed spin labelling coupled to EPR, 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 $Z_i^{\text{agg}}$ 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, this approximation is rather accurate, but under harsher conditions we do not expect this to be the case. For example, the
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. 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.

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 a new aggregation propensity profile $Z_i^{agg}$, by modulating the intrinsic aggregation propensity profile $Z_i^{agg}$, with the local stability score, $\ln P_i$.

$$Z_i^{agg} = Z_i^{agg} \left(1 - \frac{\ln P_i}{15}\right) \quad (12)$$

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 $Z_i^{agg}$ 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.

$$Z^{agg} = \frac{\sum_i Z_i^{agg} \beta(Z_i^{agg})}{\sum_i \beta(Z_i^{agg})} \quad (13)$$

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-Scheinker disease (GSS) and fatal familial insomnia (FFI). The key event in the pathogenesis of these human diseases is the conversion of the normal $\alpha$-helical protease-sensitive cellular form of the prion protein (hPrP(C)) into a $\beta$-rich form (hPrP(Sc)) that possesses distinct features such as protease resistance, insolubility and toxicity. 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. 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 $Z_i^{agg}$ profile. The similarity of the $Z_i^{agg}$ and $Z_i^{agg}$ profiles for residues 23–125 is in agreement with the experimental observation that this region is not structured. 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 $Z_i^{agg} > 1$, 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...
Therefore the comparison of the has been destabilized.

the stabilization of the hPrPSc forms after the hPrP C form suggests that the region of residues 175–193 is involved in

the secondary structure elements present in hPrP C are indicated as blue bars (β-strands) and red bars (α-helices), and the position of the disulfide bond C179–C214 is indicated by a blue line. An experimentally-determined aggregation-prone fragment (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 (Zi agg > 1) in the hPrP C form. The region corresponding to the structural core of the amyloid fibril as determined by hydrogen–deuterium exchange49 (residues 169–213) is indicated by a black bar, and corresponds to the region of high intrinsic aggregation propensity (Zi agg > 1) formed by residues 175–193.

hPrP fragments. Peptides hPrP106–114, hPrP106–126, hPrP113–126 and hPrP127–147 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 Zi 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. Therefore when the aggregation propensity profile Zi 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 experiments49 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 Zi agg. Therefore the comparison of the Zi agg and Zi agg profiles suggests that the region of residues 175–193 is involved in the stabilization of the hPrPSc 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 protofibril species,48–51 we defined a position-dependent toxicity score, Zi tox, that accounts for the propensity to form protofibrillar assemblies21

\[ Z_i^{\text{tox}} = \frac{P_i^{\text{tox}} - \mu_i^{\text{tox}}}{\sigma_i^{\text{tox}}} \]  (14)

In this equation, the terms contributing to Zi 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 Zi tox is shown in the case of Aβ 1–42 in Fig. 4.

In order to compare the predictions with the experimental results we defined an over toxicity score as21

\[ Z^{\text{tox}} = \sum_{i=1}^{N} Z_i^{\text{tox}} \]  (15)

The correlation between the Zi tox score and the toxicity of the of Aβ 1–42 mutants was found to be very high (r = 0.83) and better than the correlation obtained with the aggregation propensity score Zi agg (r = 0.75), thus supporting the

Fig. 3  Aggregation propensity profiles of the human prion protein. The black vertical bars indicate the intrinsic aggregation propensity profile, Zi agg; the red vertical bars indicate the aggregation propensity profile, Z̃i agg, calculated by taking into account the structural protection provided by the globular structure of hPrPSc form of the protein, as predicted by the lnP1 score.26,45 For reference, the secondary structure elements present in hPrP C indicated as blue bars (β-strands) and red bars (α-helices), and the position of the disulfide bond C179–C214 is indicated by a blue line. An experimentally-determined aggregation-prone fragment (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 (Zi agg > 1) in the hPrP C form. The region corresponding to the structural core of the amyloid fibril as determined by hydrogen–deuterium exchange49 (residues 169–213) is indicated by a black bar, and corresponds to the region of high intrinsic aggregation propensity (Zi agg > 1) formed by residues 175–193.

Fig. 4  Toxicity profile of Aβ 1–42. The vertical bars indicate the toxicity profile, Zi tox. The positions of the mutants whose toxicity has been studied in vivo21 are indicated by red circles. See Fig. 1 for the definition of the blue and green bars.
specific predictions are shown here for Ab oligomers (Z_{\text{agg}}), formation of \(\beta\)-rich oligomers (Z_{\text{tol}}), or formation of fibrillar aggregates (Z_{\text{agg}}). The specific predictions are shown here for Ab_{1-42}.

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.

### 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, several sequence-based methods have been proposed to achieve this goal. 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, and the method by Tartaglia et al. includes a term to describe the \(\pi\)-stacking contributions to the stability of the aggregates.

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-\(\beta\) structures, in order to assess its compatibility with this type of conformation. In the second, the propensities of polypeptide chain to self-assemble into ordered cross-\(\beta\) aggregates are estimated by constructing a knowledge-based residue-residue interaction potential using a database of native structures.

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, or according to their compatibility with the cross-\(\beta\) motif typical of ordered fibrillar assemblies, or by considering their tendency of forming \(\beta\) structures in native states. 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