doi:10.1016/j.jmb.2004.06.043 *J. Mol. Biol.* (2004) **341**, 1317–1326

JMB

Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

ELSEVIER

Prediction of the Absolute Aggregation Rates of Amyloidogenic Polypeptide Chains

Kateri F. DuBay[1], Amol P. Pawar[1], Fabrizio Chiti[2], Jesús Zurdo[1] Christopher M. Dobson[1]* and Michele Vendruscolo[1]*

[1]*Department of Chemistry*
University of Cambridge
Lensfield Road, Cambridge
CB2 1EW, UK

[2]*Dipartimento di Scienze*
Biochimiche, Viale Morgagni
50, Universitá degli Studi di
Firenze, 50134 Firenze, Italy

**Corresponding authors*

Protein aggregation is associated with a variety of pathological conditions, including Alzheimer's and Creutzfeldt-Jakob diseases and type II diabetes. Such degenerative disorders result from the conversion of the normal soluble state of specific proteins into aggregated states that can ultimately form the characteristic amyloid fibrils found in diseased tissue. Under appropriate conditions it appears that many, perhaps all, proteins can be converted *in vitro* into amyloid fibrils. The aggregation propensities of different polypeptide chains have, however, been observed to vary substantially. Here, we describe an approach that uses the knowledge of the amino acid sequence and of the experimental conditions to reproduce, with a correlation coefficient of 0.92 and over five orders of magnitude, the *in vitro* aggregation rates of a wide range of unstructured peptides and proteins. These results indicate that the formation of protein aggregates can be rationalised to a considerable extent in terms of simple physico-chemical parameters that describe the properties of polypeptide chains and their environment.

Keywords: amyloid fibrils; aggregation rates; sequence analysis; hydrophobic patterns; misfolding diseases

Introduction

Pathological conditions such as type II diabetes and neurodegenerative disorders such as Alzheimer's and Creutzfeldt-Jakob diseases have been linked with the deposition in tissue of insoluble protein aggregates.[1–5] These deposits, often in the form of amyloid plaques, are largely composed of misfolded proteins that assemble to form extended fibrillar structures.[6] Despite the lack of detectable similarities among the amino acid sequences or the native structures of amyloidogenic proteins, amyloid fibrils from different sources share common morphological and structural features.[7] Electron and atomic force microscopy have shown that amyloid fibrils are formed from protofilaments that associate laterally or twist together to form fibrils of larger diameter.[6,8–10] Moreover, amyloid fibrils show a common cross-β pattern in which the polypeptide chains form β-strands oriented perpendicularly to the long axis of the fibril, resulting in β-sheets propagating in the direction of the fibril.[7] Although amyloid deposits were initially discovered in association with several human disorders, it has recently become apparent that a wide range of other proteins, unrelated to any known disease, can form amyloid structures *in vitro* when incubated under appropriate conditions.[4,11–13] As a consequence it has been suggested that the ability to form amyloid fibrils is a common characteristic of polypeptide chains, although the ease with which they form varies greatly with the sequence.[11,14,15]

Given the increasing number of diseases that are recognised to be related to amyloid formation, and the apparent generic ability of natural and synthetic polypeptide chains to form amyloid fibrils, it is important to understand the determinants of this process. Diverse factors, both intrinsic and extrinsic to the proteins, have been reported to influence the rate of aggregation of amyloidogenic peptides and proteins. Extrinsic factors that affect the formation of protein aggregates include the interaction with cellular components such as

Present address: Kateri F. DuBay, Department of Chemistry, UC Berkeley, 419 Latimer Hall, Berkeley, CA 94720-1460, USA.

Abbreviations used: AcP, acylphosphatase.

E-mail addresses of the corresponding authors: mv245@cam.ac.uk; cmd44@cam.ac.uk

molecular chaperones,[16] proteases that generate or process the amyloidogenic precursors,[17] and the effectiveness of quality control mechanisms, such as the ubiquitin–proteasome system.[18,19] They also include physico-chemical parameters defining the environment of the polypeptides, such as pH, temperature, ionic strength and concentration.[20–25] Intrinsic factors associated with amyloid formation include a range of characteristics of polypeptide chains, such as charge,[26–29] hydrophobicity,[30–32] patterns of polar and non-polar residues,[33] and the propensities to adopt diverse secondary structure motifs.[27,32,34,35] In the case of globular proteins, the propensity to form amyloid structures is often inversely related to the stability of the native state.[36–40] Many of the proteins associated with amyloid diseases are, however, at least partially unstructured under physiological conditions. For instance, mutations in the α-synuclein gene, linked to familial forms of Parkinson's disease, cannot be correlated to alterations in protein stability, as α-synuclein appears to be natively unfolded.[41,42] Moreover, in many cases changes in stability are not sufficient to account for the pathogenic nature of mutant proteins, such as in the case of the prion protein where a substantial fraction of the mutations analysed so far have been found to have little or no effect on the stability of the native state.[43] Similarly, in familial amyloid polyneuropathies associated with transthyretin deposition, the degree of destabilisation of the native state caused by a mutation and the severity of the related clinical condition are not fully correlated.[44] It is therefore becoming clear that, even for globular proteins, intrinsic factors other than the stability of the native state must play a role in determining the propensity of a given sequence to aggregate.

The role of intrinsic properties in determining changes in the aggregation rate resulting from single amino acid substitutions has been recently analysed in detail using human muscle acylphosphatase (AcP).[45] When aggregation was studied from a denatured ensemble, a very high correlation was found between the observed changes in the aggregation rates resulting from single amino acid substitutions and the effect that each of the substitutions has on three intrinsic properties of the polypeptide chain, hydrophobicity, charge, and the propensity of the polypeptide chain to adopt α-helical or β-sheet structure.[45] 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 be influencing aggregation[32] and for peptides and proteins that are at least partially unfolded. The predicted variations in aggregation rates obtained by applying this equation to different AcP mutants showed a very good agreement ($r = 0.76$, $p < 0.001$) with the experimental results obtained from the AcP mutants.[32,45] The formula also reproduces to a remarkable extent ($r = 0.85$, $p < 0.001$) the changes in the aggregation rates observed experimentally for single amino acid substitutions in other polypeptides, including those associated with amyloid disease.[45]

Here, we take a significant step forward in this type of analysis by showing that intrinsic and extrinsic characteristics can be used as variables in a relatively simple formula that predicts accurately the absolute aggregation rates of polypeptide chains under a wide range of experimental conditions, without the requirement of experimental knowledge of the specific regions of the sequence that are particularly sensitive for aggregation. We introduce the following phenomenological equation to describe the absolute rate at which a polypeptide chain aggregates to form amyloid fibrils or their precursors:

$$\log(k) = \alpha_0 + \alpha_{\mathrm{hydr}} I^{\mathrm{hydr}} + \alpha_{\mathrm{pat}} I^{\mathrm{pat}} + \alpha_{\mathrm{ch}} I^{\mathrm{ch}} + \alpha_{\mathrm{pH}} E^{\mathrm{pH}} + \alpha_{\mathrm{ionic}} E^{\mathrm{ionic}} + \alpha_{\mathrm{conc}} E^{\mathrm{conc}} \quad (1)$$

where $\log(k)$ is the logarithm in base 10 of the aggregation rate k, in units of s^{-1}. Factors intrinsic to the amino acid sequence are denoted as I, while extrinsic, condition-dependent, factors are denoted as E. I^{hydr} represents the hydrophobicity of the sequence, calculated as the sum of the hydrophobic contributions of each residue, normalised by N, the number of amino acid residues in the sequence; the Roseman scale of hydrophobicity was used to estimate these propensities at neutral pH, using the data from Cowan and co-workers to adjust the changes in hydrophobicity experienced by amino acid residues at different pH values.[46,47] I^{pat} takes into account the existence of patterns of alternating hydrophobic–hydrophilic residues; a factor of $+1$ was assigned for each pattern of five consecutive alternating hydrophobic and hydrophilic residues in the sequence.[48] I^{ch} is the absolute value of the net charge of the sequence. E^{pH} accounts for the pH of the solution in which aggregation occurs and E^{ionic} defines the ionic strength of the solution, given in millimolar units. Finally, E^{conc} refers to the polypeptide concentration C (in millimolar units) in the solution, represented here as $\log(C + 1)$, a term always positive for any value of C.

At this stage of our investigation we focus on analysing the absolute rates of aggregation for polypeptide chains determined from their denatured states *in vitro*. We therefore exclude at the present time parameters linked to the presence of cellular components such as chaperones, proteases, and quality control systems as well as those related to the conformational stability of the folded proteins, although such factors could be included in extensions of the present study. The results of this work demonstrate our ability to correlate, and hence to predict, over a broad range of potential experimental conditions, the aggregation rates of a number of non-homologous partially unstructured peptides and proteins.

Results and Discussion

Prediction of the aggregation rates

In order to determine the coefficients in equation (1), we considered a comprehensive database from the experimental data on aggregation rates available in the literature at the time of preparation of the manuscript (see Methods). The dataset comprised data from mutational studies on the aggregation of one protein, human muscle acylphosphatase (AcP), for which extensive measurements have been reported,[28,32] and data on other systems obtained from a systematic literature search[44,49–57] (see Table 1). Aggregation rates for AcP and for transthyretin variants were determined under conditions that promote the conversion of the native state into an ensemble of unfolded or partially unfolded conformations. This permitted factors favouring aggregation to be examined in the absence of complications from changes in the stability of the native state that might occur as a consequence of the mutations. Since the remaining sequences are all peptides that do not fold into a defined globular structure, we can use experimental kinetic data while remaining confident that the changes in aggregation rates as a result of mutation are not due to changes in the stability of the globular structure.

The coefficients in equation (1) were determined by using a standard regression analysis to reproduce the experimental log(k) values for the polypeptides reported in Table 1 (see Methods). The values reported in Table 2 represent the resulting estimates of these parameters. The comparison between the calculated and observed aggregation rates for various sequences is illustrated in Figure 1; AcP variants are shown in blue and the remaining sequences in green. The linear correlation coefficient between the calculated and the observed values for the entire dataset is 0.92 ($p <$ 0.0001). The root-mean-squared error between the calculated and observed log(k) values is 0.3; this value is an estimate of the statistical error on the prediction of log(k) consistent with the results obtained by the bootstrapping test (see below).

Table 2. Results from the regression analysis of the entire dataset, which correspond to the best fit of the coefficients in equation (1)

	α	p-value
Hydrophobicity	−1.56 ± 0.38	<0.001
Patterns	0.41 ± 0.04	<0.001
Charge	−0.16 ± 0.02	<0.001
pH	0.04 ± 0.07	0.53
Ionic strength	−0.011 ± 0.001	<0.001
Concentration	0. 4 ± 0.6	0.57
Intercept	−3.3 ± 0.4	<0.001

Errors are estimated from the variability of the coefficients resulting from the bootstrap and the jackknife validation procedures; the respective p-statistics indicate the significance in the predictions, with a p-value <0.05 indicating a significant result (see the text for comments on the values for pH and concentration).

Validation of the predictions

The results of procedures such as the one described in the previous section can be highly dependent on the database used in the fitting procedure, a problem known as "overfitting". In the specific case that we have studied, an extensive testing of the approach is particularly difficult at the present stage owing to the relative scarcity of good-quality experimental data on polypeptide aggregation rates; indeed one of the objectives of the present work is to promote the measurement of the aggregation rates of a much larger number of polypeptides and proteins. Therefore, in order to test the accuracy and predictive power of equation (1) in determining the aggregation rate of any polypeptide chain we used two cross-validation methods, the bootstrapping procedure,[58] and the jackknife method.[59]

Table 1. Experimental data used in the present work

Peptide/protein	Variants	pH	Ionic strength (mM)	[Peptide] (mM)	References
AcP	59	5.5	43	0.04	28,32,61
Aβ40	2	7.4	150	0.25	53
Aβ40	1	7.4	81	0.03	55
Aβ42	1	7.4	81	0.01	55
ABri	1	9.0	89	1.31	56
AChE peptide 586–599	1	7.0	7.7	0.20	57
Amylin 22–29	2	7.2	1.1	2.0	54
Amylin 1–37	1	7.3	1.4	0.14	52
Amylin 9–37	1	7.3	1.4	0.14	52
HypF-N	1	5.5	40	0.08	75
Amglin	1	5.0	0.1	0.001	50
Leucine-rich repeats	1	7.8	3.3	0.39	49
PrP peptide 106–126	3	5.0	1.2	0.33	51
Transthyretin	3	4.4	130	0.014	44
Human calcitonin	1	7.4	25	1.5	(S. Fowler & J.Z., unpublished results)

The sequences are denoted by the common abbreviation of the peptide or protein in each case. The number of mutations whose aggregation rates were measured is given along with the experimental conditions. The sequences listed were used to fit the parameters in equation (1), with the exception of HypF-N and calcitonin that were used as a test of its predictive ability.

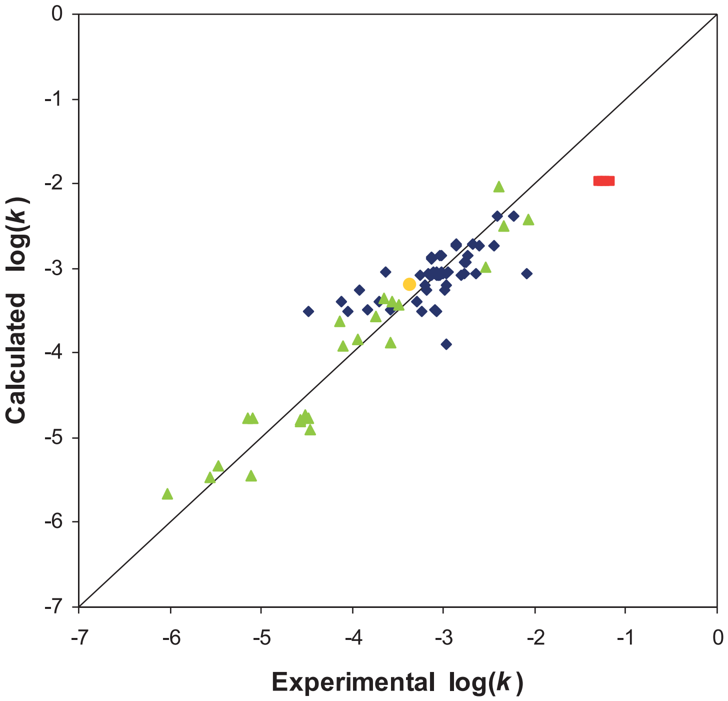

Figure 1. Results from the regression analysis of the dataset reported in Table 1. The calculated values for log(k), determined using equation (1) and the coefficients reported in Table 2, are plotted against the experimental values. Data for wild-type AcP and its mutants are shown in blue (diamonds), while data for other sequences in the dataset are shown in green (triangles). The comparison between the predicted and the experimental aggregation rates for the N-terminal domain of HypF-N is plotted as a red bar and yellow circle for human calcitonin. A line of slope 1.0 is plotted for comparison.

In the bootstrapping test, we randomly divided the entire dataset into two subsets. The first subset, composed of two-thirds of the sequences, was used as the training set, from which the coefficients in equation (1) were estimated. These coefficients were then used to predict the aggregation rates of the remaining sequences, representing the test set. The procedure was repeated 25 times, each time with a different random choice of the training set. The distribution of the correlation coefficients between the predicted and the experimental values was plotted for the training and test sets (Figure 2A). The correlation coefficient for the training set ranged from 0.90 to 0.94 with an average of 0.93; the p-value is lower than 0.0001 in all cases. The correlation coefficient for the test set ranged from 0.78 to 0.95 with an average value of 0.89; in only one of the 25 cases the correlation coefficient was lower than 0.80. We analysed carefully the list of data used in the training set in this case and found that the random selection procedure excluded all the data corresponding to a set of measurements relative to a given range of experimental conditions. The resulting overfitting of the coefficients was thus responsible for the relatively poor performance in this case. This type of effect should become less pronounced as more data on aggregation rates become available.

We then adopted the jackknife cross-validation method, in which the aggregation rate for each sequence is predicted in turn after having left that particular sequence aside during the determination of the optimal coefficients for the remaining sequences.[59] We performed this procedure for all of the wild-type and mutated polypeptides reported in Table 1. The linear correlation coefficient between predicted and observed rates was 0.91 in this case. The results of this test for various non-homologous polypeptide sequences are shown in Figure 2B. The good agreement between the predicted and experimental aggregation rates for the various proteins and peptides examined in this study shows the reliability of the formula in determining absolute aggregation rates from unstructured polypeptide chains.

Two further compelling tests for our formula are the predictions of the aggregation rate of the N-terminal domain of prokaryotic globular protein HypF (HypF-N) and human calcitonin, two proteins that were not included in the analysis described so far. The 91 residue polypeptide chain of HypF-N has been shown to form amyloid fibrils under conditions similar to those used in the AcP studies.[28,32,60] HypF-N forms amyloid fibrils even more rapidly than AcP, which has one of the fastest amyloid aggregation rates in the dataset used.[60] We predict $\log(k) = -2.0$ for HypF-N using equation (1). An experimental bound for the rate of aggregation[61] is $\log(k) \geq -1.3$. The comparison between predicted and observed aggregation rates of HypF-N (see Figure 1) shows that both values are faster than any other rate in our dataset. Similarly, the rate of amyloid aggregation for human calcitonin was found to be $\log(k) = -3.4$ at 25 mM ionic strength (S. Fowler & J.Z., unpublished results). Equation (1) predicts $\log(k) = -3.2$ for this sequence (see Figure 1).

Influence of individual factors

The values of the coefficients in equation (1) obtained in the present analysis have been

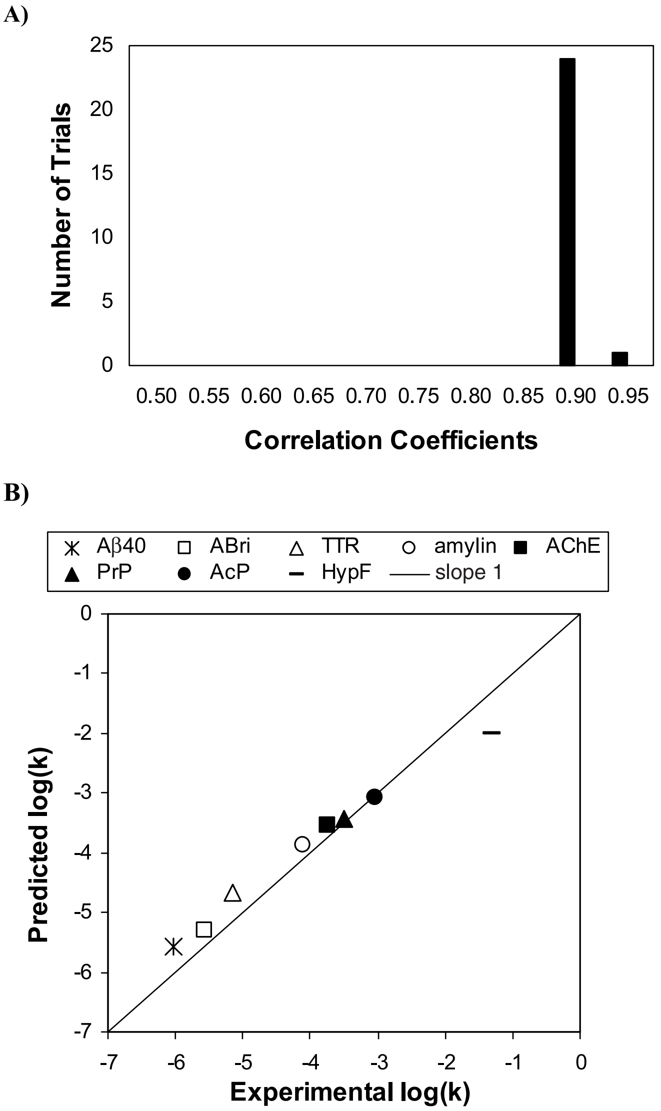

Figure 2. A, Results from the bootstrapping test for equation (1) (see Methods). The histogram shows the distribution of the correlation coefficients of both training (black) and test (grey) sets for 25 trials. B, log(*k*) values predicted for all the non-homologous wild-type sequences in our dataset by means of the jackknife cross-validation analysis. Predicted values of log(*k*) for each of the wild-type sequences shown were calculated using a regression analysis on the data for all the sequences in the dataset except those for the single wild-type sequence predicted; namely, Aβ40,[53] ABri,[56] transthyretin,[44]amylin,[52] AChE,[57] PrP,[51] AcP and HypF-N.[75] The experimental conditions for each observed aggregation rate are reported in Table 1 as well as in the references cited.

examined to discuss the extent to which they correlate individually with the factors that are considered to influence the propensity of a polypeptide chain to aggregate to form amyloid structures. We note, however, that equation (1) is phenomenological and may involve double-counting of some factors. Therefore, the interpretation of the coefficients of individual terms should be made with caution.

Intrinsic factors

Hydrophobicity

Hydrophobic interactions have long been suggested to play a significant role in amyloid formation.[62] The hydrophobicity scale that we use here assigns positive values to hydrophilic residues and negative values to hydrophobic residues;[46,63] the use of other hydrophobicity scales is also possible, requiring only a re-fitting of the coefficients in Table 2. As we found a significant ($p < 0.001$) coefficient for I^{hydr}, our analysis confirms the well-documented effect that an increased hydrophobicity leads to increased aggregation,[61] as also shown by the fact that natively unfolded proteins tend to have a low average hydrophobic content.[64]

Hydrophobic patterns

Hydrophobic patterning was found in the present study to be one of the most significant ($p <$ 0.001) determinants of aggregation rates in equation (1). The importance of hydrophobic–hydrophilic patterns has been extensively studied by Hecht and co-workers,[33,65–68] and alternating patterns of the type that we used have been shown to be among the least common features of natural protein sequences.[48] A length of five consecutive hydrophobic and hydrophilic alternating residues was found to yield the most significant correlation with the experimental values of aggregation kinetics. The positive value of the coefficient for patterns indicates that the more patterns of this type occur in a given sequence, the faster is its aggregation rate. Although a previous study[31] observed a selection against consecutive hydrophobic residues in natural sequences, we detected no correlation in our dataset between aggregation rates and either consecutive hydrophobic or consecutive hydrophilic residues.

Charge

The highly significant ($p < 0.001$) negative coefficient of the charge contribution indicates that the aggregation rate of a polypeptide chain is inversely proportional to the absolute value of the net charge; such a correlation has been observed before for AcP and its mutants.[28] In the case of a small peptide, however, charges of ±1 were shown to be more favourable to the formation of highly regular amyloid structures than net charges[29] of 0 or ±2. These studies suggest, however, that rapid aggregation may be associated with the formation of poorly defined structures, whereas the formation of highly ordered amyloid seems to require longer periods of incubation.[29] Modification of the functional form of I^{ch} from a linear to a polynomial expression with maxima at ±1 gave a lower correlation coefficient for the equation. Thus, our investigation supports a linear dependence of the aggregation kinetics on the absolute value of the net charge of the polypeptide.[45]

Extrinsic factors

pH

The coefficients for pH and for concentration (see below) could not be fitted precisely, owing to the limited size of the available database on aggregation rates. Nevertheless, we included them in equation (1) because we anticipate that the measurement of additional aggregation rates will allow better estimates to be obtained in the future. Our present results suggest that the pH may be positively correlated to the rate of aggregation. This finding is apparently not consistent with the observation that formation of amyloid fibrils often occurs at low pH.[11,12,24,34,69] A possible explanation is that the charge is included explicitly as an intrinsic factor. The increased positive charge that proteins tend to have at low pH is expected to disfavour aggregation, but such an effect is more than counterbalanced by the lowering of the stability of most proteins with decreasing pH.

Ionic strength

We found that taking into consideration the ionic strength improved the accuracy of the predictions of the aggregation rates. The extension of the dataset to include aggregation rates measured for the same peptide or protein under a wider range of salt concentrations than the one we studied here (from 0.1 mM to 150 mM) will enable a more accurate rationalization of this effect.

Peptide concentration

Our results indicate that, as expected, the rate of aggregation may increase with the peptide concentration, C, although a fully quantitative analysis of this effect is not possible at present (see above). Several authors have proposed the existence of a critical concentration for amyloid formation, which is specific for each particular system.[70] However, since all the experimental data that we considered were obtained above the critical concentration required for aggregation, the extrapolation of equation (1) to very low concentrations should be considered with considerable caution.

Additional factors

Factors, such as temperature, stirring and native-state stability, are known to influence amyloid aggregation rates significantly. High temperatures are generally found to lead to faster aggregation rates[20,23] but the limited range of temperatures in the available experimental results included in the dataset (298–310 K) makes it difficult to establish reliably its contribution at the present time. Similarly, although the extent to which solutions are agitated is known to influence greatly the kinetics of aggregation, this factor was not considered here (see Methods). Lastly, as already mentioned, the present procedure does not take into account the stability of the native state, but rather predicts rates of aggregation from unfolded states. For globular proteins such states are likely to be populated at low levels under physiological conditions; nevertheless, they may still determine the observed aggregation rates.[71] In principle, however, the stability of the native state, which is likely to be an important factor in determining aggregation rates in many cases, could also be considered as an additional factor in the formula, once a sufficiently large set of experimental data becomes available.

Although β-sheet and α-helical propensities have been found to influence significantly the changes in amyloid formation rates with mutations in a variety of studies,[34,35,45,72] they were not found to be statistically significant in our analysis and therefore were not included in equation (1). Secondary structure propensities are known, however, to be relevant in describing aggregation rates for regions directly involved in the aggregation process,[45] or when secondary structure propensities do not lead to the formation of stable secondary structural motifs.[34] In addition, Hecht and co-workers[33,65–67] showed that amino acid patterns are a major determinant in defining the secondary structure adopted by polypeptides, and in influencing the ability of amino acid sequences to form amyloid structures. The inclusion of patterns in equation (1) might, therefore account for some of the effects previously attributed to secondary structure propensities.[34,35,45]

The present analysis has been carried out using experimental data available in the literature at the time the manuscript was prepared (see Methods). The relative scarcity of these data has prevented at this stage the consideration of the effects of some additional factors, such as those mentioned above, known to influence amyloid aggregation rates. For a similar reason, as additional data become available, estimates for parameters in the ranges for which we have optimized the coefficients in equation (1) should also become more reliable. The methodology that we present is, however, general, and we hope that the opportunity to refine the quantitative analysis will promote the systematic collection of data relevant to these additional factors in order to rationalize their effects.

Conclusions

We have shown that relatively simple parameters defining a polypeptide sequence and its environment can determine, under a wide range of experimental conditions, its aggregation rate from an unfolded or partially unfolded state into amyloid fibrils. In the present work we have analysed the effects of intrinsic properties of the polypeptide sequence, such as hydrophobicity, hydrophobic–hydrophilic patterning and charge, and also environmental parameters, such as pH, ionic strength and concentration. Other factors, such as stability of the native state, temperature and the degree to which the solution is agitated could also be included in the formula, if suitable data become available to enable a reliable determination of their coefficients.

The aggregation rates calculated using the approach that we present here reproduce the experimentally observed rates with a correlation coefficient of 0.92 (bootstrap cross-validated 0.89, jackknife cross-validated 0.91). They can, therefore, be expected to allow accurate predictions within the ranges of conditions included in the dataset used at the present time, namely protein lengths from 8 to 127 residues, pH values from 4.4 to 9.0, ionic strengths from 0.1 mM to 150 mM, temperatures from 298 K to 310 K and peptide concentrations from 0.001 mM to 2 mM. We should also note that the formula derived in this work was obtained by neglecting the observation that some regions of a polypeptide chain are likely to be more important than others for determining the aggregation rates.[32] The approach presented here considers the overall properties of an amino acid sequence, with all residues having the same relative importance. This approximation is likely to be responsible for the negligible influence of secondary structure propensities that was found in the present analysis. When such propensities were analysed using the experimental knowledge of the regions important for the aggregation of AcP, the results demonstrated their importance for the prediction of changes in aggregation rates caused by single amino acid substitution.[45] However, even with the limitations imposed by neglecting the existence of these regions, perhaps compensated for by the inclusion of hydrophobic patterns, we have found a robust correlation between predicted and experimental aggregation rates for essentially any sequence of residues. The quality of the predictions presented here is likely to be improved further by combining equation (1) with an algorithm capable of predicting the relative importance of different regions in a polypeptide chain. At the same time, the fact that the regions important for aggregation do not need to be experimentally determined in order to use this formula enhances greatly its general applicability.

The present analysis is applicable to the kinetic behaviour measured after the lag phase in the aggregation, which is a common feature of aggregation resulting in highly organised amyloid fibrils, as indeed for crystallization. After the lag phase, single-exponential behaviour is generally observed. We hope that the encouraging results we have presented here will stimulate experimental groups, in addition to our own, to carry out systematic studies of the factors that influence the duration of the lag phase as well as the subsequent growth phase, thus gathering a body of knowledge that will make it possible to rationalise its origin and the factors on which it depends.

One of the most important conclusions of the present work is that, under the conditions of applicability of equation (1) that we have discussed, a relatively small number of physico-chemical parameters of a polypeptide chain and its environment can be used to determine its intrinsic propensity to form amyloid aggregates, with no apparent influence of the mechanism of aggregation and structure adopted by the polypeptide chain in the resulting aggregates. In addition to providing a computational tool for determining *a priori* the rate of a process with so many implications in protein science, biotechnology and medicine, this finding supports further the suggestion that protein aggregation is a generic process where the common backbone of a polypeptide chain plays a dominant role, although amino acid side-chains modulate the propensity of the backbone to aggregate as well as many details of the resulting structure. The ability to predict the aggregation propensity of a given peptide or protein with the accuracy shown here for a range of rates varying by a factor of 10^5 should be a powerful tool to assist experimental studies of the behaviour of natural polypeptides and their propensity to aggregate, as well as to establish the principles by which sequences have been selected through evolution to avoid misfolding and aggregation. A quantitative understanding of the factors influencing aggregation rates will increase our capability to predict the onset of amyloidoses and other protein deposition diseases, in addition to helping us explore effective therapeutic strategies. It will also help us to design or modify polypeptides and proteins rationally, to enhance their properties of folding and self-association for biotechnology, pharmaceutical developments and structural biology.

Methods

Dataset

Kinetic data on the aggregation of AcP and its mutants were obtained from the literature;[28,32] in these studies thioflavin T fluorescence was used to determine the rate of aggregation of each protein in solution. All AcP data were measured under identical conditions and provided the largest set of data used in the present analysis (60 sequences). The second set of data included the aggregation rates of a range of peptides under different conditions, obtained from published results (see Table 1). A literature search was conducted using "kinetics" and "fibril" or "amyloid" as keywords, resulting in an initial list of over 800 references. We then selected those studies that described quantitative measurements of the rates of aggregation of short peptides or denatured proteins in a buffer solution that formed fibrils detectable by electron microscopy over the course of the experiment. Because of the difficulties in quantifying different stirring procedures in a variety of sample vessels, no aggregation experiments performed under stirring or agitation were considered. This selection procedure led us to ten references that provided kinetic data on 23 sequences under various salt conditions, occasionally with small amounts of co-solvent remaining from the peptide stock solution. Sequences were chosen using the criteria described above, prior to any kinetic analysis. No sequences were excluded after the analysis, nor were new ones added.

Aggregation rates were determined from kinetic traces obtained by the following methods: thioflavin T fluorescence, turbidity, CD, or direct estimation of the relative amount of aggregated material using techniques such as sedimentation, size-exclusion chromatography, and filtration. Although these methods detect slightly different aspects of aggregation, they are closely linked, and in some systems where two or more experimental techniques have been applied, a similar kinetic profile has been observed.[32,52,57] The values of log(k) determined by different methods in these papers differ by less than 0.1 unit[32,52] in all but one case,[57] where turbidity kinetics and thioflavin T kinetics differ by 0.8 unit, perhaps as a result of other differences in experimental procedure. In the experimental studies that we considered, mass/volume analyses were used in the absence of an independent technique to confirm the results. However, since these methods may be considered the most direct method of observing the growth of physical aggregates, the data obtained solely by these methods were included in the analysis.

Lag phases were not considered in our analysis, as they are often not reported or difficult to extract from the published data. While a comprehensive understanding of lag phases in protein aggregation is still lacking, they appear to be particularly susceptible to slight changes in aggregation procedure, as illustrated by various studies where seeded and non-seeded solutions result in nearly identical elongation rates.[50,73] Thus, the present analysis focuses on the aggregation kinetics after the lag phase where an elongation phase with single exponential behaviour is generally observed. Kinetic traces were fitted to the equation $y = A(1 - e^{-kt})$ where k is the rate constant in units of s^{-1}. The logarithm in base 10 of the rate constant, log(k), was used in equation (1), since the values of log(k) were better described by a normal distribution than the value of k itself.

Derivation and validation of the formula

The functional form of each factor in equation (1) was chosen after examining a variety of phenomenological combinations of the factors likely to influence the propensity to aggregate. We considered two classes of factors, intrinsic and extrinsic. Intrinsic factors included properties of the amino acid sequence, such as hydrophobicity, hydrophobic patterns and charge. Their functional forms were determined by examining a subset of AcP mutants to find the representation that best correlated with changes in log(k) amongst the mutants. The extrinsic factors included peptide concentration, ionic strength, and pH. We used a logarithm form for the term describing the effect of the peptide concentration in order to avoid overestimating rates at higher concentrations. This choice is also supported by recent observations of the critical nucleus of aggregation, where the dependence of log(k) on log(C) has been found to be linear.[74] All other terms considered here were assumed to be approximately linear.

Regressions were performed with the software Rweb 1.8.0† to obtain the coefficients in equation (1) that minimize the differences between the calculated and experimental log(k) values. In interpreting the meaning of the numerical constants in the formula we should note again its phenomenological nature. The formula may contain double-counting for some factors (e.g. hydrophobicity and hydrophobi, patterns, pH and charge); this is not problematic, as the coefficients are fitted from experimental data and not derived from first principles. The formula was validated using both bootstrap and jackknife cross-validation techniques as described in the text.[58,59]

Acknowledgements

We are grateful for support from the Gates Cambridge Trust (to K.F.D. and A.P.P.), the Leverhulme Trust (to J.Z., C.M.D. and M.V.), the Italian MIUR and CNR (to F.C.), the Wellcome Trust (to C.M.D.) and the Royal Society (to M.V.).

References

1. Horwich, A. L. & Weissman, J. S. (1997). *Cell*, **89**, 499–510.
2. Kelly, J. W. (1998). *Curr. Opin. Struct. Biol.* **8**, 101–106.
3. Rochet, J. C. & Lansbury, P. T. (2000). *Curr. Opin. Struct. Biol.* **10**, 60–68.

† http://www.math.montana.edu/Rweb/

4. Dobson, C. M. (2001). *Phil. Trans. Roy. Soc. ser. B*, **356**, 133–145.
5. Selkoe, D. J. (2002). *Science*, **298**, 789–791.
6. Sunde, M. & Blake, C. (1997). *Advances in Protein Chemistry*, vol. 50. pp. 123–159.
7. Sunde, M., Serpell, L. C., Bartlam, M., Fraser, P. E., Pepys, M. B. & Blake, C. C. F. (1997). *J. Mol. Biol.* **273**, 729–739.
8. Jimenez, J. L., Guijarro, J. L., Orlova, E., Zurdo, J., Dobson, C. M., Sunde, M. & Saibil, H. R. (1999). *EMBO J.* **18**, 815–821.
9. Harper, J. D., Wong, S. S., Lieber, C. M. & Lansbury, P. T. (1999). *Biochemistry*, **38**, 8972–8980.
10. Serpell, L. C., Sunde, M., Benson, M. D., Tennent, G. A., Pepys, M. B. & Fraser, P. E. (2000). *J. Mol. Biol.* **300**, 1033–1039.
11. Guijarro, J. I., Sunde, M., Jones, J. A., Campbell, I. D. & Dobson, C. M. (1998). *Proc. Natl Acad. Sci. USA*, **95**, 4224–4228.
12. Chiti, F., Webster, P., Taddei, N., Clark, A., Stefani, M., Ramponi, G. & Dobson, C. M. (1999). *Proc. Natl Acad. Sci. USA*, **96**, 3590–3594.
13. Dobson, C. M. (2002). *Nature*, **418**, 729–730.
14. Dobson, C. M. (1999). *Trends Biochem. Sci.* **24**, 329–332.
15. Stefani, M. & Dobson, C. M. (2003). *J. Mol. Med.* **81**, 678–699.
16. Muchowski, P. J. (2002). *Neuron*, **35**, 9–12.
17. Citron, M., Westaway, D., Xia, W. M., Carlson, G., Diehl, T., Levesque, G. *et al.* (1997). *Nature Med.* **3**, 67–72.
18. Bence, N. F., Sampat, R. M. & Kopito, R. R. (2001). *Science*, **292**, 1552–1555.
19. Tofaris, G. K., Razzaq, A., Ghetti, B., Lilley, K. S. & Spillantini, M. G. (2003). *J. Biol. Chem.* **278**, 44405–44411.
20. Abrahamson, M. & Grubb, A. (1994). *Proc. Natl Acad. Sci. USA*, **91**, 1416–1420.
21. Lomakin, A., Chung, D. S., Benedek, G. B., Kirschner, D. A. & Teplow, D. B. (1996). *Proc. Natl Acad. Sci. USA*, **93**, 1125–1129.
22. Harper, J. D. & Lansbury, P. T. (1997). *Annu. Rev. Biochem.* **66**, 385–407.
23. Kusumoto, Y., Lomakin, A., Teplow, D. B. & Benedek, G. B. (1998). *Proc. Natl Acad. Sci. USA*, **95**, 12277–12282.
24. Su, Y. & Chang, P. T. (2001). *Brain Res.* **893**, 287–291.
25. Zurdo, J., Guijarro, J. I., Jimenez, J. L., Saibil, H. R. & Dobson, C. M. (2001). *J. Mol. Biol.* **311**, 325–340.
26. Konno, T. (2001). *Biochemistry*, **40**, 2148–2154.
27. Tjernberg, L., Hosia, W., Bark, N., Thyberg, J. & Johansson, J. (2002). *J. Biol. Chem.* **277**, 43243–43246.
28. Chiti, F., Calamai, M., Taddei, N., Stefani, M., Ramponi, G. & Dobson, C. M. (2002). *Proc. Natl Acad. Sci. USA*, **99**, 16419–16426.
29. de la Paz, M. L., Goldie, K., Zurdo, J., Lacroix, E., Dobson, C. M., Hoenger, A. & Serrano, L. (2002). *Proc. Natl Acad. Sci. USA*, **99**, 16052–16057.
30. Otzen, D. E., Kristensen, O. & Oliveberg, M. (2000). *Proc. Natl Acad. Sci. USA*, **97**, 9907–9912.
31. Schwartz, R., Istrail, S. & King, J. (2001). *Protein Sci.* **10**, 1023–1031.
32. Chiti, F., Taddei, N., Baroni, F., Capanni, C., Stefani, M., Ramponi, G. & Dobson, C. M. (2002). *Nature Struct. Biol.* **9**, 137–143.
33. West, M. W., Wang, W. X., Patterson, J., Mancias, J. D., Beasley, J. R. & Hecht, M. H. (1999). *Proc. Natl Acad. Sci. USA*, **96**, 11211–11216.
34. Villegas, V., Zurdo, J., Filimonov, V. V., Aviles, F. X., Dobson, C. M. & Serrano, L. (2000). *Protein Sci.* **9**, 1700–1708.
35. Kallberg, Y., Gustafsson, M., Persson, B., Thyberg, J. & Johansson, J. (2001). *J. Biol. Chem.* **276**, 12945–12950.
36. Hurle, M. R., Helms, L. R., Li, L., Chan, W. N. & Wetzel, R. (1994). *Proc. Natl Acad. Sci. USA*, **91**, 5446–5450.
37. Quintas, A., Saraiva, M. J. M. & Brito, R. M. M. (1999). *J. Biol. Chem.* **274**, 32943–32949.
38. Chiti, F., Taddei, N., Bucciantini, M., White, P., Ramponi, G. & Dobson, C. M. (2000). *EMBO J.* **19**, 1441–1449.
39. Ramirez-Alvarado, M., Merkel, J. S. & Regan, L. (2000). *Proc. Natl Acad. Sci. USA*, **97**, 8979–8984.
40. Siepen, J. A. & Westhead, D. R. (2002). *Protein Sci.* **11**, 1862–1866.
41. Weinreb, P. H., Zhen, W. G., Poon, A. W., Conway, K. A. & Lansbury, P. T. (1996). *Biochemistry*, **35**, 13709–13715.
42. Li, J., Uversky, V. N. & Fink, A. L. (2001). *Biochemistry*, **40**, 11604–11613.
43. Liemann, S. & Glockshuber, R. (1999). *Biochemistry*, **38**, 3258–3267.
44. Hammarstrom, P., Jiang, X., Hurshman, A. R., Powers, E. T. & Kelly, J. W. (2002). *Proc. Natl Acad. Sci. USA*, **99**, 16427–16432.
45. Chiti, F., Stefani, M., Taddei, N., Ramponi, G. & Dobson, C. M. (2003). *Nature*, **424**, 805–808.
46. Roseman, M. A. (1988). *J. Mol. Biol.* **200**, 513–522.
47. Cowan, R. & Witthaker, R. G. (1990). *Peptide Res.* **3**, 75–80.
48. Broome, B. M. & Hecht, M. H. (2000). *J. Mol. Biol.* **296**, 961–968.
49. Symmons, M. F., Buchanan, S. G. S., Clarke, D. T., Jones, G. & Gay, N. J. (1997). *FEBS Letters*, **412**, 397–403.
50. Kayed, R., Bernhagen, J., Greenfield, N., Sweimeh, K., Brunner, H., Voelter, W. & Kapurniotu, A. (1999). *J. Mol. Biol.* **287**, 781–796.
51. Salmona, M., Malesani, P., De Gioia, L., Gorla, S., Bruschi, M., Molinari, A. *et al.* (1999). *Biochem. J.* **342**, 207–214.
52. Goldsbury, C., Goldie, K., Pellaud, J., Seelig, J., Frey, P., Muller, S. A. *et al.* (2000). *J. Struct. Biol.* **130**, 352–362.
53. Miravalle, L., Tokuda, T., Chiarle, R., Giaccone, G., Bugiani, O., Tagliavini, F. *et al.* (2000). *J. Biol. Chem.* **275**, 27110–27116.
54. Azriel, R. & Gazit, E. (2001). *J. Biol. Chem.* **276**, 34156–34161.
55. Fezoui, Y. & Teplow, D. B. (2002). *J. Biol. Chem.* **277**, 36948–36954.
56. El-Agnaf, O. M. A., Sheridan, J. M., Sidera, C., Siligardi, G., Hussain, R., Haris, P. I. & Austen, B. M. (2001). *Biochemistry*, **40**, 3449–3457.
57. Cottingham, M. G., Hollinshead, M. S. & Vaux, D. J. T. (2002). *Biochemistry*, **41**, 13539–13547.
58. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (2002). *Numerical Recipes in C++*, Cambridge University Press, Cambridge.
59. Mardia, K. W., Kent, J. T. & Bibby, J. M. (1979). *Multivariate Analysis*, Academic Press, London.
60. Chiti, F., Taddei, N., Stefani, M., Dobson, C. M. & Ramponi, G. (2001). *Protein Sci.* **10**, 879–886.
61. Calamai, M., Taddei, N., Stefani, M., Ramponi, G. & Chiti, F. (2003). *Biochemistry*, **42**, 15078–15083.
62. Fink, A. L. (1998). *Fold. Des.* **3**, R9–R23.

63. Creighton, T. E. (1993). *Proteins. Structure and Molecular Properties*, W. H. Freeman, Co., New York.
64. Uversky, V. N. (2003). *Cell. Mol. Life Sci.* **60**, 1852–1871.
65. Kamtekar, S., Schiffer, J. M., Xiong, H. Y., Babik, J. M. & Hecht, M. H. (1993). *Science*, **262**, 1680–1685.
66. West, M. W. & Hecht, M. H. (1995). *Protein Sci.* **4**, 2032–2039.
67. Xiong, H. Y., Buckwalter, B. L., Shieh, H. M. & Hecht, M. H. (1995). *Proc. Natl Acad. Sci. USA*, **92**, 6349–6353.
68. Wang, W. X. & Hecht, M. H. (2002). *Proc. Natl Acad. Sci. USA*, **99**, 2760–2765.
69. Bouchard, M., Zurdo, J., Nettleton, E. J., Dobson, C. M. & Robinson, C. V. (2000). *Protein Sci.* **9**, 1960–1967.
70. Lomakin, A., Teplow, D. B., Kirschner, D. A. & Benedek, G. B. (1997). *Proc. Natl Acad. Sci. USA*, **94**, 7942–7947.
71. Booth, D. R., Sunde, M., Bellotti, V., Robinson, C. V., Hutchinson, W. L., Fraser, P. E. *et al.* (1997). *Nature*, **385**, 787–793.
72. Wood, S. J., Wetzel, R., Martin, J. D. & Hurle, M. R. (1995). *Biochemistry*, **34**, 724–730.
73. Padrick, S. B. & Miranker, A. D. (2002). *Biochemistry*, **41**, 4694–4703.
74. Thakur, A. K. & Wetzel, R. (2002). *Proc. Natl Acad. Sci. USA*, **99**, 17014–17019.
75. Chiti, F., Bucciantini, M., Capanni, C., Taddei, N., Dobson, C. M. & Stefani, M. (2001). *Protein Sci.* **10**, 2541–2547.

Edited by A. R. Fersht

(Received 19 December 2003; received in revised form 4 June 2004; accepted 6 June 2004)