doi:10.1016/j.jmb.2010.08.013 J. Mol. Biol. (2010) 402, 919–928

Contents lists available at www.sciencedirect.com

Journal of Molecular Biology

journal homepage: http://ees.elsevier.com.jmb

Proteome-Level Interplay between Folding and Aggregation Propensities of Proteins

Gian Gaetano Tartaglia* and Michele Vendruscolo*

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

Received 30 March 2010; received in revised form 5 August 2010; accepted 9 August 2010
Available online 13 August 2010

Edited by M. Sternberg

Keywords:
protein folding;
protein aggregation;
CamFold;
Zyggregator

With the advent of proteomics, there is an increasing need of tools for predicting the properties of large numbers of proteins by using the information provided by their amino acid sequences, even in the absence of the knowledge of their structures. One of the most important types of predictions concerns whether proteins will fold or aggregate. Here, we study the competition between these two processes by analyzing the relationship between the folding and aggregation propensity profiles for the human and *Escherichia coli* proteomes. These profiles are calculated, respectively, using the CamFold method, which we introduce in this work, and the Zyggregator method. Our results indicate that the kinetic behavior of proteins is, to a large extent, determined by the interplay between regions of low folding and high aggregation propensities.

Introduction

The amino acid sequences of proteins play a crucial role in determining their folding behavior.[1,2] Indeed, substantial progress has been made in the prediction of native-state structures of proteins by using only the information provided by their sequences, so that it is currently possible to generate models for the structures of small globular proteins with a relatively high degree of confidence and accuracy, as shown by the increasing quality of the results of the CASP exercise.[3] It has also been shown that from the knowledge of the amino acid sequence of a protein it is possible to predict its folding rate;[4–8] in principle, the folding pathway of a protein should be predictable by just considering its sequence, at least after deriving from the sequence itself a model of the native state topology.[9,10]

It has also been established that the amino acid sequences of proteins determine, to a large extent, also their aggregation behavior.[11–18] One particularly interesting aspect of such predictions is that it is possible to differentiate the aggregation propensity from the unfolded state, or "intrinsic aggregation propensity",[13] from that from the folded state, or "structure-corrected aggregation propensity".[16] The latter type of aggregation propensity takes into account the fact that, in the folded state, regions that are highly aggregation prone are protected from aggregation because they are buried within the native structure and hence not exposed to the solved and unavailable to form intermolecular interactions.

Consistent with these ideas, several recent studies have suggested that there should be a competition between folding and aggregation,[16,19–24] in the sense that, in order to avoid aggregation, the regions of the amino acid sequence that are highly aggregation prone should be protected during the folding process and in the folded state. In order to investigate this competition, we introduce in this work the concept of "folding propensity profile" of amino acid sequences. We define this property in terms of the physicochemical properties of the amino acids—the folding propensity of a given region of a polypeptide sequence is defined in terms of its hydrophobicity, secondary-structure propensity, and electrostatic charge. The resulting Z_i^{fold} score gives the folding

*Corresponding authors. E-mail addresses: gian.tartaglia@crg.es; mv245@cam.ac.uk.

Present address: G. G. Tartaglia, Gene Function and Evolution, Centre for Genomic Regulation (CRG), UPF, Barcelona, Spain.

Abbreviation used: AcP, acylphosphatase.

0022-2836/$ - see front matter © 2010 Published by Elsevier Ltd.

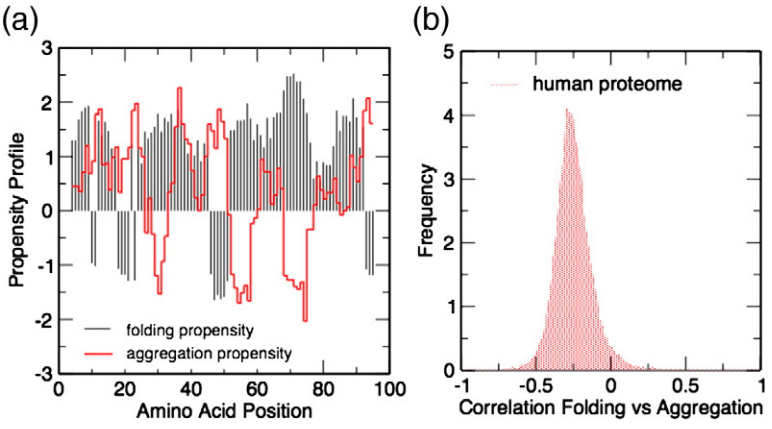

Fig. 1. Relationship between folding and aggregation propensity profiles. (a) Comparison between Z_i^{agg}, the aggregation propensity profile, and Z_i^{fold}, the folding propensity profile, for AcP; an anti-correlation (coefficient of correlation, −0.5) is observed between the two profiles. (b) Relationship between folding and aggregation propensity profiles in the human proteome. The distribution of the coefficients of correlation between the Z_i^{agg} and the Z_i^{fold} profiles has a mean of 0.3 and a variance of 0.1. These findings indicate that the rate-limiting regions for folding tend to have high aggregation propensities.

propensity as a function of the residue number along the amino acid sequence. The parameters used to define the Z_i^{fold} profiles are determined by comparing the experimental folding rates for a set of proteins with the corresponding overall Z_{fold} scores, which are obtained by summing the Z_i^{fold} scores over all amino acid sequences.

Our analysis of the human proteome indicates that the folding propensity profiles are anticorrelated with the intrinsic aggregation propensity profiles and that at the same time the intrinsic aggregation propensity scores are anticorrelated with the structure-corrected aggregation propensity scores. These results suggest that the kinetic behavior of proteins is to a large extent determined by the balance between regions that are rate-determining for folding and have high intrinsic aggregation propensities.

Results

A relationship between folding and intrinsic aggregation propensities

The folding propensity profile of a given amino acid sequence is defined in this work using its physicochemical properties, including hydrophobicity, secondary-structure propensity, and electrostatic charge (see Methods). Folding propensity profiles are calculated by using the CamFold method† (see Methods), which provides them in terms of Z_i^{fold} scores.

Since the overall folding rate of a protein is defined as the sum over the sequence of the Z_i^{fold} scores of individual amino acids (see Methods), regions of low Z_i^{fold} scores tend to lower the folding rate, and therefore, they are expected to correspond to the rate-limiting steps of the folding process. We illustrate this point by considering the regions that play an important role in the folding process of acylphosphatase (AcP). The folding kinetics of AcP has been studied by protein engineering methods, and the key residues for folding have been identified as Y11, P54, and F94;[25,26] residues are considered key for folding if they are required for defining the topology of the native state.[26] The folding propensity profile, which was calculated using Eqs. (1)–(3) (see Methods), exhibits minima in correspondence of these regions (Fig. 1a).

Further, in order to address the main goal of this work, which is to analyze the relationship between folding and aggregation propensities of proteins, we compared the folding propensity profile of AcP with its aggregation propensity profile. The aggregation propensity profiles were calculated by using the Zyggregator method,[16,27] which was introduced to identify the regions that play a major role in the aggregation process. The comparison of the folding and the aggregation propensity profiles in the case of AcP suggests that regions of low folding propensity tend to have a high aggregation propensity (Fig. 1a). Similar results are found in a series of additional examples (see Supplementary Data, Figs. S1–S3).

To confirm the insight provided by the analysis of these initial cases, we carried out a proteome-level analysis of the correlation between folding and

Table 1. Summary of the propensity scores for the transitions between the native, unfolded, and fibrillar states

	Native	Unfolded
Native	—	Z_{UN} (Z_{fold})
Unfolded	Z_{NU} (Z_{unfold})	—
Fibril	Z_{NF} (Z_{agg}^{SC})	Z_{UF} (Z_{agg})

In this work, we defined a propensity score to fold, that is, to go from the unfolded (U) to the native state (N), Z_{UN}, or Z_{fold}. Previously, we defined the propensity score to go from the unfolded to the fibrillar state (F),[13,28] Z_{UF}, or Z_{agg}, as well as the propensity score to go from the folded to the fibrillar state,[16,27] Z_{NF}, or Z_{agg}^{SC} (structure-corrected Z_{agg} score). In principle, it is also possible to define the propensity score to unfold, Z_{NU}, or Z_{unfold}, that is, to go from the native to the unfolded state.

† http://www-vendruscolo.ch.cam.ac.uk/camfold.php

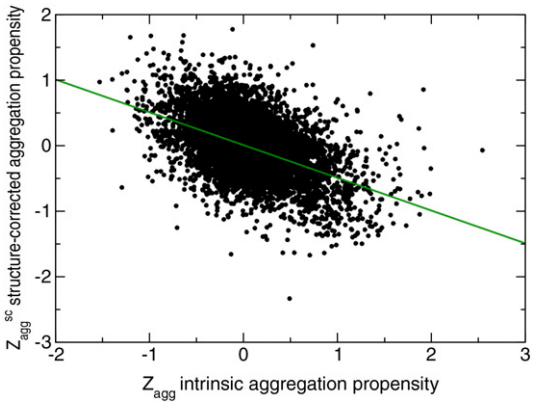

Fig. 2. Relationship between the Z_{agg} and Z_{agg}^{SC} aggregation propensity scores. The intrinsic aggregation propensity (i.e., the tendency to go from the unfolded to the fibrillar state, Z_{UF}, or Z_{agg}) is anticorrelated with the aggregation propensity with structural corrections (i.e., the tendency to go from the folded to the fibrillar state, Z_{NF}, or Z_{agg}^{SC}). Each point represents a protein in the human proteome; the coefficient of correlation is −0.5.

aggregation propensity profiles, which indicates the presence of a significant anticorrelation between these two profiles (Fig. 1b). The presence of a detectable anticorrelation at the proteome level is particularly interesting in the view that folding and aggregation are likely to be encoded differently in the amino acid sequences of proteins, the former at a global level, while the latter at a local level. As noted above, our current understanding indicates that the aggregation process is often driven by the formation of intermolecular interactions involving specific regions of the amino acid sequence of a protein.[11–18] By contrast, the folding process, at least within individual protein domains, is instead often more cooperative and involves the participation of the whole polypeptide chain. In terms of propensity profiles, therefore, it would seem much more straightforward to define an aggregation propensity profile to identify aggregation-prone and aggregation-resistant regions, than a folding propensity profile, a difficult task that requires to single out fast-folding and slow-folding regions in sequences that behave to a large extent cooperatively. We have shown (Fig. 1a) that in the case of AcP, a protein that folds in a nucleation–condensation mechanism, it is possible to characterize the regions that play a key role in the rate-limiting step as having low folding propensity scores. Such regions could involve the specific amino acids that form the folding nucleus, as in AcP or TI I27 (see Supplementary Data, Fig. S1), or structural motifs in which the amino acids that form the folding nucleus are located, as in CI2 (see Supplementary Data, Fig. S2). More in general, however, particularly for proteins that fold in a modular manner rather than cooperatively (see

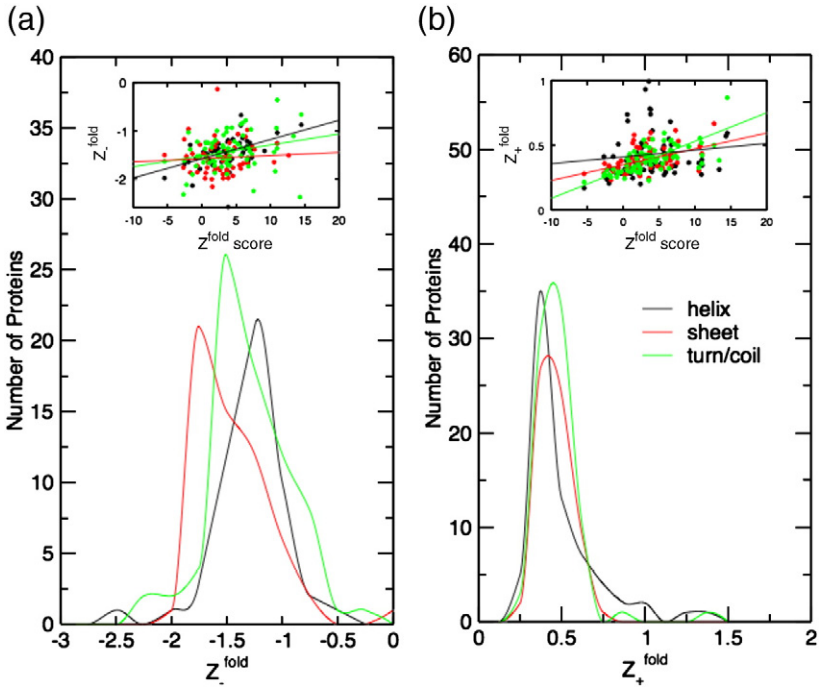

Fig. 3. Structural characterization of the contributions to the Z_{fold} score. Histograms of the Z_-^{fold} (a) and Z_+^{fold} (b) values calculated separately for different secondary-structure elements: α-helices (black), β-sheets (red), and turns/coils (green). The correlations of the overall Z_{fold} scores with the Z_-^{fold} and Z_+^{fold} scores for different secondary-structure elements are shown in the insets. (a) β-Sheet regions give the largest contributions to the Z_-^{fold} scores. β-Sheet and turn/coil regions present broad distributions of Z_-^{fold} values and have coefficient of correlations of, respectively, 0.10 and 0.24 with the overall Z_{fold} scores (inset). By contrast α-helical elements show a narrower distibution of Z_-^{fold} values and a coefficient of correlation of 0.47 with the overall Z_{fold} scores (inset). (b) The distributions of Z_+^{fold} values for α-helical, β-sheet, and turn/coil regions present similar widths as the case of Z_-^{fold} values. Coil/turn elements are associated with a high content of polar amino acids and present a coefficient of correlation with Z_{fold} scores of 0.64 (inset). Z_+^{fold} values for β-sheet elements and overall Z_{fold} scores show a coefficient of correlation of 0.44, while Z_+^{fold} values for α-helical elements present a broader distribution and a coefficient of correlation of 0.1 (inset). The STRIDE algorithm[29] was used to assign secondary-structure elements using the three-dimensional structures.

Table 2. Analysis of the amino acid composition of the regions of the sequence that give positive and negative contributions to the Z_{fold} score, that is, of high and low folding propensity, respectively

Amino acid type	Composition (scale 1)[a]	<0 (scale 2)[b]	>0 (scale 3)[c]
A	0.96	0.64	0.36
C	0.00	0.36	0.64
D	0.64	0.41	0.59
E	0.96	0.57	0.43
F	0.34	0.62	0.38
G	0.91	0.51	0.49
H	0.15	0.38	0.62
I	0.59	0.62	0.38
K	0.96	0.65	0.35
L	1.00	0.39	0.61
M	0.11	0.29	0.71
N	0.39	0.26	0.74
P	0.38	0.00	1.00
Q	0.34	0.68	0.32
R	0.48	0.48	0.52
S	0.59	0.55	0.45
T	0.63	0.55	0.45
V	0.82	0.77	0.23
W	0.07	1.00	0.00
Y	0.28	0.83	0.17

[a] Average amino acid composition score (scale 1) for the proteins used to validate the algorithm; compositions are normalized from 0 to 1for each amino acid type, so that L and C are the most and less frequent amino acid types in the database, respectively (Table S1).

[b] Negative Z_{fold} contribution score [scale 2, see Eq. (7)]; amino acids that give the most negative contributions to the overall Z_{fold} score tend to have high hydrophobicity[30] and β-sheet propensity[31] (see also Table 3); scores are normalized between 0 and 1 using the normalized amino acid composition (scale 1), so that W and P give the largest and smallest negative contributions, respectively, to the overall Z_{fold} score.

[c] Positive Z_{fold} contribution score [scale 3, see Eq. (8)]; amino acids that give the most positive contributions to the overall Z_{fold} score tend to have high polarity[37] and turn/coil propensity[36] (see also Table 3); as in the case of negative Z_{fold} contributions, scores are normalized between 0 and 1 using the normalized amino acid composition (scale 1), so that P and W give the largest and smallest positive contributions, respectively, to the overall Z_{fold} score.

Supplementary Data, Fig. S4), or that form macromolecular complexes after folding (see Supplementary Data, Fig. S2), low folding propensity scores might also be associated with other molecular processes, such as the docking of preformed domains.

Taken together, the results presented in this section suggest that, as a general trend at the proteome level, regions that are rate limiting in the folding process tend also to promote aggregation, thus indicating that regions that are important for determining the folding process are also important for determining the aggregation process.

A relationship between folding and structure-corrected aggregation propensities

In order to understand why under normal conditions proteins fold instead of aggregating, given that similar regions play an important role in promoting both processes, we analyzed the aggregation propensity with structural corrections,[16,27] that is, the aggregation propensity profiles in the native state. For clarity, we first summarize the three types of propensities that we consider in this work (Table 1): first, the propensity score to fold, Z_{UN}, or Z_{fold}, that is, to go from the unfolded (U) to the native state (N); second, the propensity score to convert from the unfolded to the fibrillar state (F),[13,28] Z_{UF}, or Z_{agg}; third, the propensity score to convert from the folded to the fibrillar state,[16,27] Z_{NF}, or Z_{agg}^{SC} (structure-corrected Z_{agg} score). We also mention that, in principle, it should also be possible to formulate a method of calculating the propensity score to unfold, Z_{NU}, or Z_{unfold}, that is, to go from the native to the unfolded state.

We found that the intrinsic aggregation propensity (i.e., from the unfolded to the fibrillar state, Z_{UF} or Z_{agg}) is anticorrelated with the structure-corrected aggregation propensity (i.e., from the folded to the fibrillar state, Z_{NF} or Z_{agg}^{SC}) (Fig. 2). These results, taken together with that discussed in the previous section that folding propensity is anticorrelated with the intrinsic aggregation propensity, indicate that the regions that play an important role in the folding process have also a high aggregation propensity, but this propensity tends to be suppressed by the folding process itself.

Regions of low folding propensity tend to have low disorder and high β-sheet propensities

In this section, we characterize the contributions from different secondary-structure elements to the overall Z_{fold} scores. We first distinguish negative (Z_{-}^{fold}) and positive (Z_{+}^{fold}) contributions to the Z_{fold} scores for different secondary-structure elements [see Methods, Eqs. (5) and (6)]. We found that β-sheet regions tend to provide the largest contributions to Z_{-}^{fold} scores (Fig. 3). More in general, our results indicate that negative contributions to the Z_{fold} score are associated with a high content of residues with high hydrophobicity and β-sheet propensity, while positive contributions to the Z_{fold} score are associated

Table 3. Characterization of the contributions to the overall Z_{fold} scores of the different amino acid types in terms of their β-sheet, aggregation, and disorder propensities

	Scale 1	Scale 2	Scale 3
β-Sheet	0.64[31]	0.55[32]	0.50[33]
Aggregation	0.45[13]	0.40[34]	0.32[12]
Disorder	−0.60[35]	−0.56[36]	−0.47[32]

The amino acid scale of negative contributions to Z_{fold} (scale 2, see Table 2) correlates with the β-sheet (coefficient of correlation of 0.55) and aggregation propensity scales (coefficient of correlation of 0.40) and anticorrelates with the disorder propensity scale (coefficient of correlation of −0.56). Opposite trends are found for positive contributions to Z_{fold} (scale 3, see Table 2).

with a high content of residues with high polarity and turn/coil propensity (Table 2).

The characterization of the regions that give positive and negative contributions to the overall Z_{fold} scores in terms of their β-sheet, aggregation, and disorder propensities indicates that the rate-determining regions for folding, which tend to have low Z_{fold} scores, have also a low disorder propensity score and a high aggregation propensity score (Table 3). Our results thus indicate that the rate-limiting regions for folding are characterized by a low propensity to be disordered and a significant aggregation propensity. By contrast, low-aggregation propensity regions tend to have high folding propensity scores. We also observe that negative Z_{fold} regions are less abundant than Z_{fold} positive regions, which might result as a consequence of a selective pressure acting to reduce the aggregation potential of proteins.

A relationship between folding, aggregation, and GroEL/ES requirements

In this section, we present an analysis of the role played by molecular chaperones in the competition between aggregation and folding. We analyzed a set of 250 *Escherichia coli* proteins whose GroEL/ES requirements were determined experimentally.[38] These proteins were divided in three classes, according to their level of dependence on GroEL/ES, which was deduced from *in vitro* and *in vivo* refolding assays: (i) class I, in which proteins fold largely independently of GroEL/ES but may use it to optimize their folding yield; (ii) class II, in which substrates are highly chaperone dependent, at least under mildly unfavorable environmental conditions, but can utilize either DnaK/DnaJ or GroEL/ES for

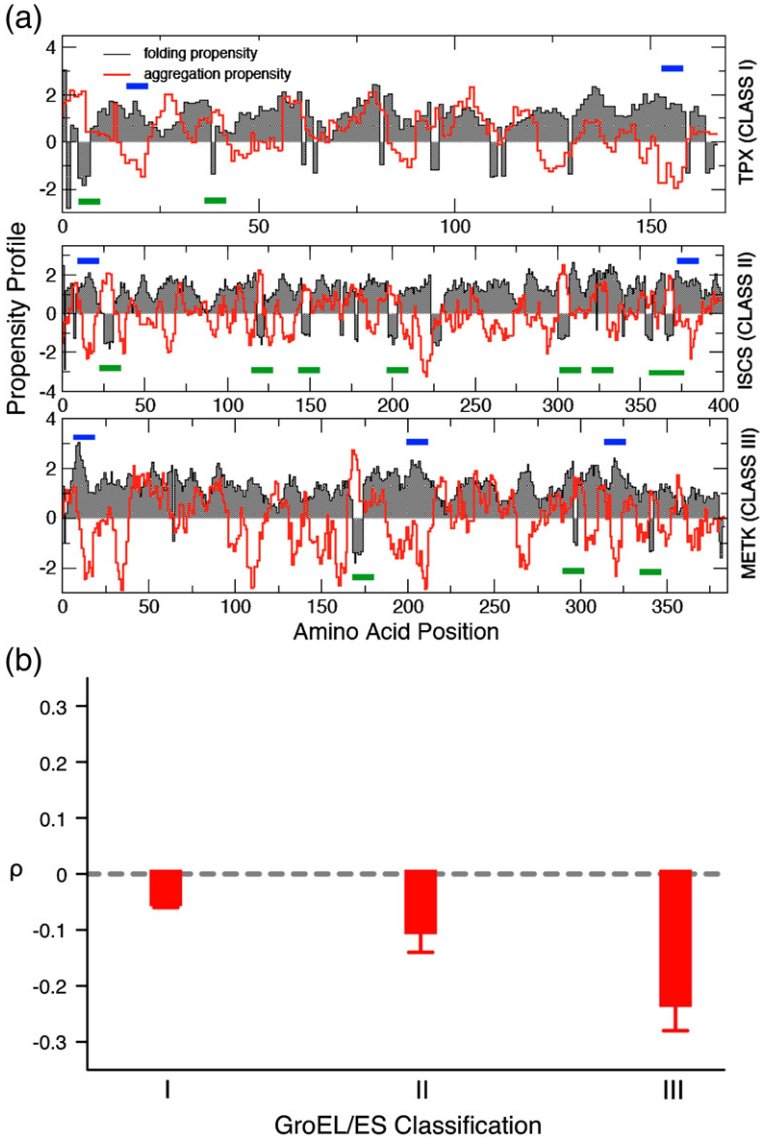

Fig. 4. (a) Comparison of the aggregation propensity profiles for three *E. coli* proteins representative of the GroEL interaction classes[38]: TPX (class I), ISCS (class II), and METK (class III). (b) Average values for the *E. coli* proteins in the three GroEL interaction classes of the coefficient of correlation, ρ, between the folding and the aggregation propensity profiles.

folding; and (iii) class III, in which substrates have an absolute requirement for the GroEL/EL system in order to fold correctly.

By considering initially three proteins representative of each GroEL/ES class (Fig. 4a), we identified regions of negative Z_{fold} and positive Z_{agg} scores (green segments in Fig. 4a, which indicate significant correlation between the two values), as well as the regions of positive Z_{fold} and negative Z_{agg} scores (blue segments in Fig. 4a). The negative peaks of Z_{fold} associated with the positive peaks of Z_{agg} correspond to regions that remain exposed for a long time during the folding process despite having a high aggregation propensity. The analysis of these three proteins suggests that the regions of low Z_{fold} and high Z_{agg} scores (green regions in Fig. 4a) tend to be more abundant in proteins of classes II and III. In order to generalize this result, we analyzed the correlation over the entire set of 250 proteins whose GroEL/ES class was determined experimentally, finding the anticorrelation between Z_{fold} and Z_{agg} scores to be more significant in proteins of classes II and III (Fig. 4b), consistent with the idea that the presence of regions of high aggregation and low folding propensities tend to reduce the reliability of the folding process and to increase the level of dependence on molecular chaperones.

Discussion

In this work, we have studied the relationship between the folding and aggregation propensities of the different regions of amino acid sequences. In order to perform sequence-based predictions of the folding behavior of proteins, we have introduced the CamFold method, which enables the calculation of the folding propensity profiles of proteins by using physicochemical properties of the different regions of their amino acid sequences. We have also used the same type of properties to obtain sequence-based predictions of the aggregation behavior of proteins, which we performed using the Zyggregator method.[16] As both the CamFold and the Zyggregator calculations are very fast, these methods are particularly suitable for proteome-level studies.

We performed an analysis of the human and *E. coli* proteomes, finding that the rate-determining regions for folding also tend to have a high aggregation propensity. We also described, however, how such regions do not actually promote aggregation under normal cellular conditions because, as the folding process is faster than the aggregation process, they become buried in the native state before they can form stable intermolecular interactions.

Methods

Definition of the folding propensity of an amino acid sequence

We define first a scale of intrinsic folding propensities of individual amino acids on the basis of their physicochemical properties. The intrinsic folding propensity of an amino acid of type a (Table 4) is defined as

$$p^{fold}(a) = \alpha_h h(a)\alpha_s s(a)\alpha_t t(a)\alpha_H H(a)\alpha_C C(a) \qquad (1)$$

where $h(a)$, $s(a)$, and $t(a)$ are the secondary-structure propensities (α-helix, β-sheet, and turn, respectively); $H(a)$ is the hydrophobicity; and $C(a)$ is the electrostatic charge. For each physicochemical property, we use a consensus of experimentally determined scales. Four scales were used to estimate the contribution of hydrophobicity,[39–42] six for the β-sheet propensity,[43] eight for the α-helical propensity,[32,44] and three for the propensity to form turns[33,36,45] (Table 4). As most of the scales are correlated (the average coefficient of correlation is 0.41),[43,44] the number of effectively independent

Table 4. Folding propensity and physicochemical scales at pH 7

AA	Z_{fold}	Hydrophobicity	Turn	α-Helix	β-Sheet
A	0.34	0.69	0.22	0.57	0.34
C	0.97	0.8	0.70	0.62	0.23
D	0.61	0.17	0.74	0.67	0.35
E	0.73	0.43	0.32	0.63	0.14
F	0.31	0.75	0.20	0.56	0.92
G	0.58	0.55	0.95	0.00	0.13
H	0.47	0.48	0.56	0.48	0.14
I	0.55	0.97	0.00	0.45	0.99
K	0.50	0.00	0.60	0.87	0.07
L	0.65	1.00	0.21	0.48	0.87
M	0.55	0.95	0.21	0.32	0.40
N	0.48	0.35	1.00	0.32	0.90
P	1.00	0.88	0.66	0.44	0.00
Q	0.35	0.47	0.47	0.51	0.72
R	0.63	0.34	0.46	1.00	0.22
S	0.73	0.39	0.79	0.45	0.06
T	0.44	0.49	0.34	0.30	0.14
V	0.55	0.95	0.09	0.45	0.78
W	0.00	0.49	0.40	0.00	0.62
Y	0.21	0.43	0.50	0.54	1.00

The scales for hydrophobicity, turn, α-helix, and β-strand propensities were obtained from a consensus of experimentally determined propensities.[32,33,36,39–45] We used a linear combination of these scales to determine the Z_{fold} scale. The β-sheet propensity is associated with a negative weight with the Z_{fold} scale, while hydrophobicity and charge and the turn and α-helix propensities have positive weights. The individual coefficients of correlation with the Z_{fold} scale are −0.49 for the β-strand propensity, 0.30 for the α-helix propensity, 0.26 for the turn propensity, and 0.22 for hydrophobicity. The hydrophobicity scale used here has a correlation of 0.85 with the hydrophobicity scale determined by Cowan and Whittaker at pH 7;[39] the turn propensity has a coefficient of correlation of 0.90 with the Deleage and Roux turn scale;[36] the α-helix and β-strand propensities show coefficients of correlation of 0.55 and 0.64, with the respective scales determined by Chou and Fasman.[33] All the scales are normalized between 0 and 1.

parameters of our model is estimated to be of about 12. In addition, the intrinsic folding propensity score is pH dependent, because all the scales that we used are pH dependent; hence, our model could be used also to predict the folding rate at different pH values.

In order to take account of the effect of the local sequence context on the intrinsic folding propensities, we calculated a position-dependent folding propensity score as

$$P_i^{\text{fold}} = \frac{1}{7}\sum_{j=-3}^{3} p_{i+j}^{\text{fold}} + \alpha_{\text{aro}} I_i^{\text{aro}} + \alpha_{\text{pat}} I_i^{\text{pat}} \qquad (2)$$

where $p_i^{\text{fold}} = p^{\text{fold}}(a_i)$ is the intrinsic propensity of the amino acid a at position i. We considered the average of the intrinsic folding propensities of a seven-residue segment of the protein centered at position i. The window size for the average was reduced for the segments in proximity of the N-terminus and C-terminus. The terms and account for the presence, respectively, of polar/non-polar[46] and aromatic[47] patterns and are defined to be 1 if residue i is included in a polar/non-polar or aromatic pattern and 0 otherwise. These terms are used to characterize the aggregation propensity of the unfolded aggregated state, which is in competition with the propensity to fold of the polypeptide chain.[27]

The P_i^{fold} score is normalized as[16]

$$Z_i^{\text{fold}} = \frac{P_i^{\text{fold}} - \mu^{\text{fold}}}{\sigma^{\text{fold}}} \qquad (3)$$

where μ^{fold} and σ^{fold} are the average and standard deviation of the P_i^{fold} score over a set of random amino acid sequences.

From the Z_i^{fold} score, we define an overall folding propensity Z_{fold} by summing over the contributions characterized by high Z_i^{fold} scores and subtracting the contributions of low Z_i^{fold} scores

$$Z^{\text{fold}} = \frac{\sum_{i=1}^{N} \left(Z_i^{\text{fold}}\right)^{\nu_1} \vartheta\left(Z_i^{\text{fold}} - \alpha\right)}{\left[\sum_{i=1}^{N} \vartheta\left(Z_i^{\text{fold}} - \alpha\right)\right]^{\nu_1}} L^{\nu_0} - \frac{\sum_{i=1}^{N} \left|Z_i^{\text{fold}}\right|^{\nu_2} \vartheta\left(-Z_i^{\text{fold}} + \beta\right)}{\left[\sum_{i=1}^{N} \vartheta\left(-Z_i^{\text{fold}} + \beta\right)\right]^{\nu_2}} L^{\nu_0} \qquad (4)$$

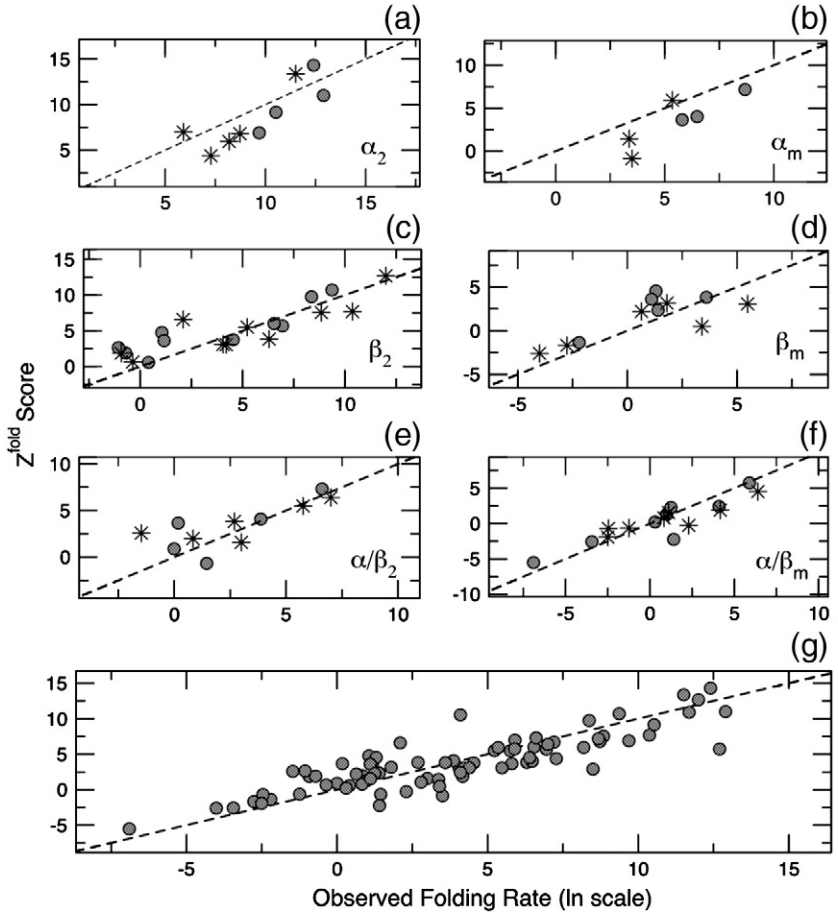

Fig. 5. Fitting of the parameters of the CamFold method. In order to determine the parameters to define the folding propensity profiles, we matched the overall Z_{fold} scores of Eq. (4) with the folding rates of 90 proteins (Supplementary Data, Table S1, and Fig. S2). The folding rates are partitioned in six groups according to their secondary structure and their two- or multi-state kinetics: (a) two-state α-proteins, (b) multi-state α-proteins, (c) two-state β-proteins, (d) multi-state β-proteins, (e) two-state α/β-proteins, and (f) multi-state α/β-proteins. Proteins used for training are marked with asterisks, while proteins employed for testing are represented with circles. To assess the statistical significance of the results, we divided each data set into two equal parts that were chosen randomly (see also Supplementary Data). For this test, 100 random partitions were generated; representative examples are reported in the plots. The individual coefficients of correlation of the six data sets are used to estimate the accuracy of the algorithm: (a) $r = 0.83$ (train) and $r = 0.80$ (test); (b) $r = 0.98$ (train) and $r = 0.92$ (test); (c) $r = 0.85$ (train) and $r = 0.88$ (test); (d) $r = 0.83$ (train) and $r = 0.84$ (test); (e) $r = 0.80$ (train) and $r = 0.77$ (test); (f) $r = 0.90$ (train) and $r = 0.91$ (test). (g) Using a leave-one-out procedure (see Methods), we calculate a correlation coefficient $r = 0.87$ for the six data sets combined together. The p value for this data set is 10^{-5}, indicating very high significance for the correlation. The unitary slope reference line is reported in all the plots.

where $\vartheta(Z_i^{fold}-\alpha)$ defines the region characterized by high Z_i^{fold}, $\vartheta(-Z_i^{fold}+\beta)$ defines the region characterized by low Z_i^{fold}, the function $\vartheta(Z_i^{fold})$ is 1 for $Z_i^{fold}\geq 0$ and 0 for $Z_i^{fold}<0$, α and β are the thresholds, L is the chain length of the protein, and ν_0, ν_1, and ν_2 and are scaling factors.

Determination of the CamFold parameters by a fitting of known folding rates

In order to define the parameters α in Eqs. (1) and (2) for the CamFold method for calculating folding propensity profiles, we matched the overall Z_{fold} scores given by Eq. (4) with the folding rates using a database of 90 proteins available in literature[4,5,48,49] (Fig. 5, Supplementary Table S1 and Fig. S5). The coefficient of correlation between the Z_{fold} score and the experimentally measured folding rates is 0.87 (Fig. 5), which was estimated using a leave-one-out cross-validation procedure; in this method, the protein whose Z_{fold} score is to be calculated is not used to fit the parameters. The results that we obtained are comparable to those obtained using the information provided by the secondary structure, which achieves a coefficient of correlation of 0.82 between calculated and experimental folding rates for two- and three-state proteins.[5] For comparison, the coefficient of correlation between the absolute contact order[10] and the folding rates is -0.87 if two-state proteins in our database are considered and -0.74 if multi-state proteins are analyzed (see Supplementary Data).

The predictive ability provided by the CamFold method is significantly separated from the corresponding values obtained upon randomization of the experimental rates (Fig. S5). The likelihood of obtaining coefficients of correlation >0.50 with a random set of folding rates is extremely small ($p<10^{-5}$). We also considered whether the Z_{fold} scores could be used to predict the differences in folding rates for homologous proteins; the three families that we considered (α-spectrins, immunoglobulins, and AcPs) indicate that there is a high correlation between folding rates and the Z_{fold} scores in these cases (Fig. S6). In this work, we did not use the experimental folding rates available for protein mutants. As a result, the current implementation of the CamFold method is not biased towards specific protein classes that are more extensively studied in literature, but it may require modifications to predict the effects of amino acid mutations in the different protein families.

The parameters associated with electrostatic charge, α-helix, and turn propensities are positive, while β-sheet propensities present a negative contribution (Table 4). The overall Z_{fold} score increases with the ability of the polypeptide chain to form α-helices and establish van der Waals or electrostatic interactions but anticorrelates with the β-sheet content. The frequency of Z_i^{fold} negative minima correlates significantly with the β-sheet content and might represent the time delay required to establish tertiary interactions. In addition, the N- and C-termini often show negative Z_i^{fold} scores, which is a consequence of the poor ability of these regions to form strong contacts.

Negative and positive contributions to the Z_{fold} score

We define the negative Z_{fold} score of a polypeptide chain of N residues, Z_-^{fold}, by summing the negative Z_i^{fold} scores of individual amino acids

$$Z_-^{fold}=\frac{\sum_{i=1}^{N}|Z_i^{fold}|^{\nu_-}\vartheta\left(-Z_i^{fold}+\beta_-\right)}{\left[\sum_{i=1}^{N}\vartheta\left(-Z_i^{fold}+\beta_-\right)\right]^{\nu_-}} \quad (5)$$

Similarly, we define the positive Z_{fold} score, indicated by Z_+^{fold}, by summing the positive Z_i^{fold} scores for individual amino acids

$$Z_+^{fold}=\frac{\sum_{i=1}^{N}|Z_i^{fold}|^{\nu_+}\vartheta\left(Z_i^{fold}-\beta_+\right)}{\left[\sum_{i=1}^{N}\vartheta\left(Z_i^{fold}-\beta_+\right)\right]^{\nu_+}} \quad (6)$$

where β_+, β_-, ν_+, and ν_- are parameters available upon request.

Negative and positive scales in Table 2

The negative scale (scale 2) in Table 2 is calculated as

$$\text{Scale}_-=\frac{\text{Composition}\left(Z_-^{fold}\right)-\text{Composition}\left(Z_+^{fold}\right)}{\text{Composition}\left(Z_+^{fold}\right)+\text{Composition}\left(Z_-^{fold}\right)} \quad (7)$$

Similarly, the positive scale (scale 3) is calculated as

$$\text{Scale}_+=\frac{\text{Composition}\left(Z_+^{fold}\right)-\text{Composition}\left(Z_-^{fold}\right)}{\text{Composition}\left(Z_+^{fold}\right)+\text{Composition}\left(Z_-^{fold}\right)} \quad (8)$$

Software availability

The CamFold method is freely available‡, and the Zyggregator method is freely available for academic users§.

Acknowledgements

We thank Dr J. Clarke for discussions and for providing folding rates reported in Fig. S2. This work was supported by the Leverhulme Trust, the Medical Research Council, Spanish Ministry of Science, the European Molecular Biology Organization, and the Royal Society.

‡ http://www-vendruscolo.ch.cam.ac.uk/camfold.php
§ http://www-vendruscolo.ch.cam.ac.uk/zyggregator.php

Supplementary Data

Supplementary data to this article can be found online at doi:10.1016/j.jmb.2010.08.013

References

1. Anfinsen, C. B. (1973). Principles that govern folding of protein chains. *Science*, **181**, 223–230.
2. Fersht, A. R. (1999). *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*. W. H. Freeman, New York, NY.
3. Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B. & Tramontano, A. (2009). Critical assessment of methods of protein structure prediction—Round VIII. *Proteins*, **77**, 1–4.
4. Gromiha, M. M. (2005). A statistical model for predicting protein folding rates from amino acid sequence with structural class information. *J. Chem. Inf. Model.* **45**, 494–501.
5. Ivankov, D. N. & Finkelstein, A. V. (2004). Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. *Proc. Natl Acad. Sci. USA*, **101**, 8942–8944.
6. Gromiha, M. M., Thangakani, A. M. & Selvaraj, S. (2006). Fold-rate: prediction of protein folding rates from amino acid sequence. *Nucleic Acids Res.* **34**, W70–W74.
7. Ma, B. G., Guo, J. X. & Zhang, H. Y. (2006). Direct correlation between proteins' folding rates and their amino acid compositions: an ab initio folding rate prediction. *Proteins*, **65**, 362–372.
8. Jiang, Y. F., Iglinski, P. & Kurgan, L. (2009). Prediction of protein folding rates from primary sequences using hybrid sequence representation. *J. Comp. Chem.* **30**, 772–783.
9. Guerois, R. & Serrano, L. (2000). The sh3-fold family: experimental evidence and prediction of variations in the folding pathways. *J. Mol. Biol.* **304**, 967–982.
10. Plaxco, K. W., Simons, K. T. & Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985–994.
11. Chiti, F., Stefani, M., Taddei, N., Ramponi, G. & Dobson, C. M. (2003). Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature*, **424**, 805–808.
12. Fernandez-Escamilla, A. M., Rousseau, F., Schymkowitz, J. & Serrano, L. (2004). Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.* **22**, 1302–1306.
13. Pawar, A. P., DuBay, K. F., Zurdo, J., Chiti, F., Vendruscolo, M. & Dobson, C. M. (2005). Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases. *J. Mol. Biol.* **350**, 379–392.
14. Conchillo-Sole, O., de Groot, N. S., Aviles, F. X., Vendrell, J., Daura, X. & Ventura, S. (2007). Aggrescan: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics*, **8**, 65.
15. Trovato, A., Seno, F. & Tosatto, S. C. E. (2007). The pasta server for protein aggregation prediction. *Protein Eng. Des. Sel.* **20**, 521–523.
16. Tartaglia, G. G., Pawar, A. P., Campioni, S., Dobson, C. M., Chiti, F. & Vendruscolo, M. (2008). Prediction of aggregation-prone regions in structured proteins. *J. Mol. Biol.* **380**, 425–436.
17. Goldschmidt, L., Teng, P. K., Riek, R. & Eisenberg, D. (2010). Identifying the amylome, proteins capable of forming amyloid-like fibrils. *Proc. Natl Acad. Sci. USA*, **107**, 3487–3492.
18. Maurer-Stroh, S., Debulpaep, M., Kuemmerer, N., de la Paz, M. L., Martins, I. C., Reumers, J. *et al.* (2010). Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat. Methods*, **7**, 237–242.
19. Chiti, F. & Dobson, C. M. (2009). Amyloid formation by globular proteins under native conditions. *Nat. Chem. Biol.* **5**, 15–22.
20. Friel, C. T., Smith, D. A., Vendruscolo, M., Gsponer, J. & Radford, S. E. (2009). The mechanism of folding of im7 reveals competition between functional and kinetic evolutionary constraints. *Nat. Struct. Mol. Biol.* **16**, 318–324.
21. Hamada, D., Tanaka, T., Tartaglia, G. G., Pawar, A., Vendruscolo, M., Kawamura, M. *et al.* (2009). Competition between folding, native-state dimerisation and amyloid aggregation in beta-lactoglobulin. *J. Mol. Biol.* **386**, 878–890.
22. Pechmann, S., Levy, E. D., Tartaglia, G. G. & Vendruscolo, M. (2009). Physicochemical principles that regulate the competition between functional and dysfunctional association of proteins. *Proc. Natl Acad. Sci. USA*, **106**, 10159–10164.
23. Castillo, V. & Ventura, S (2009). Amyloidogenic regions and interaction surfaces overlap in globular proteins related to conformational diseases. *PLoS Comp. Biol.* **5**, e1000476.
24. Tzotzos, S. & Doig, A. J. (2010). Amyloidogenic sequences in native protein structures. *Protein Sci.* **19**, 327–348.
25. Chiti, F., Taddei, N., White, P. M., Bucciantini, M., Magherini, F., Stefani, M. & Dobson, C. M. (1999). Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nat. Struct. Biol.* **6**, 1005–1009.
26. Vendruscolo, M., Paci, E., Dobson, C. M. & Karplus, M. (2001). Three key residues form a critical contact network in a protein folding transition state. *Nature*, **409**, 641–645.
27. Tartaglia, G. G. & Vendruscolo, M. (2008). The zyggregator method for predicting protein aggregation propensities. *Chem. Soc. Rev.* **37**, 1395–1401.
28. Dubay, K. F., Pawar, A. P., Chiti, F., Zurdo, J., Dobson, C. M. & Vendruscolo, M. (2004). Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains. *J. Mol. Biol.* **341**, 1317–1326.
29. Heinig, M. & Frishman, D. (2004). Stride: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res.* **32**, W500–W502.
30. Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H. & Zehfus, M. H. (1985). Hydrophobicity of amino-acid residues in globular-proteins. *Science*, **229**, 834–838.
31. Lifson, S. & Sander, C. (1979). Antiparallel and parallel beta-strands differ in amino-acid residue preferences. *Nature*, **282**, 109–111.

32. Levitt, M. (1978). Conformational preferences of amino-acids in globular proteins. *Biochemistry*, **17**, 4277–4284.
33. Chou, P. Y. & Fasman, G. D. (1978). Prediction of the secondary structure of proteins form their amino acid sequence. *Adv. Enzymol.* **47**, 45–148.
34. de Groot, N. S., Pallares, I., Aviles, F. X., Vendrell, J. & Ventura, S (2005). Prediction of "hot spots" of aggregation in disease-linked polypeptides. *BMC Struct. Biol.* **5**, 18.
35. Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J. & Russell, R. B. (2003). Protein disorder prediction: implications for structural proteomics. *Structure*, **11**, 1453–1459.
36. Deleage, G. & Roux, B. (1987). An algorithm for protein secondary structure prediction based on class prediction. *Protein Eng.* **1**, 289–294.
37. Zimmerman, J. N., Eliezer, N. & Simha, R. (1968). Characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.* **21**, 170.
38. Kerner, M. J., Naylor, D. J., Ishihama, Y., Maier, T., Chang, H. C., Stines, A. P. *et al.* (2005). Proteome-wide analysis of chaperonin-dependent protein folding in escherichia coli. *Cell*, **122**, 209–220.
39. Cowan, R. & Witthaker, R. G. (1990). Hydrophobicity indices for amino acid residues as determined by hplc. *Peptide Res.* **3**, 75–80.
40. Creighton, T. E. (1993). *Proteins. Structure and Molecular Properties.* W. H. Freeman & Co., New York, NY.
41. Meek, J. L. (1980). Prediction of peptide retention times in high-pressure liquid-chromatography on the basis of amino-acid-composition. *Proc. Natl Acad. Sci. USA*, **77**, 1632–1636.
42. Wimley, W. C. & White, S. H. (1996). Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat. Struct. Biol.* **3**, 842–848.
43. de la Paz, M. L. & Serrano, L. (2004). Sequence determinants of amyloid fibril formation. *Proc. Natl Acad. Sci. USA*, **101**, 87–92.
44. Pace, C. N. & Scholtz, J. M. (1998). A helix propensity scale based on experimental studies of peptides and proteins. *Biophys. J.* **75**, 422–427.
45. Huang, F. & Nau, W. M. (2003). A conformational flexibility scale for amino acids in peptides. *Angew. Chem., Int. Ed.* **42**, 2269–2272.
46. Xiong, H. Y., Buckwalter, B. L., Shieh, H. M. & Hecht, M. H. (1995). Periodicity of polar and nonpolar amino acids is the major determinant of secondary structure in self-assembling oligomeric peptides. *Proc. Natl Acad. Sci. USA*, **92**, 6349–6353.
47. Gazit, E. (2002). A possible role for pi-stacking in the self-assembly of amyloid fibrils. *FASEB J.* **16**, 77–83.
48. Fulton, K. F., Bate, M. A., Faux, N. G., Mahmood, K., Betts, C. & Buckle, A. M. (2007). Protein folding database (pfd 2.0): an online environment for the international foldeomics consortium. *Nucleic Acids Res.* **35**, D304–D307.
49. Zhou, H. Y. & Zhou, Y. Q. (2002). Folding rate prediction using total contact distance. *Biophys. J.* **82**, 458–463.