Structural determinants of the rate of protein folding

Bengt Nöltинг*, Wolfram Schālke, Patrick Hampel, Florian Grundig, Siegfried Gantert, Nicole Sips, Wolfhard Bandlow, Phoebe X. Qi

Prussian Private Institute of Technology at Berlin, Am Schlosspark 30, Berlin D-13187, Germany
Institute for Genetics and Microbiology, Maria-Ward-Strasse 1a, Munich D-80638, Germany
US Department of Agriculture, ARS, ERRC, 600 E. Mermaid Lane, Wyndmoor, PA 19038, USA

Received 28 June 2002; received in revised form 5 December 2002; accepted 14 February 2003

Abstract

To understand the mechanism of protein folding and to assist rational design of fast-folding, non-aggregating and stable artificial enzymes, it is essential to determine the structural parameters which govern the rate constants of folding, $k_f$. It has been found that $-\log k_f$ is a linear function of the so-called chain topology parameter ($CTP$) within the range of $10^{-1} \text{s}^{-1} \leq k_f \leq 10^{8} \text{s}^{-1}$. The correlation between $-\log k_f$ and $CTP$ is much improved than using previously published contact order ($CO$) method. It has been further suggested that short sequence separations may be preferred for the establishment of stable interactions for the design of novel artificial enzymes and the modification of slow-folding proteins with aggregating intermediates.

Keywords: Protein folding kinetics; Chain topology; $\Phi$-value analysis; Helix–coil transition; Nucleation–condensation mechanism; Rational design of enzymes

1. Introduction

There has been much studies in the field of protein folding including protein thermostability (Backmann et al., 1998; Maes and Backmann, 1999), hydrophobic (Drablos, 1999; Garcia-Hernandez and Hernandez-Araña, 1999; Chan, 2000; Czaplewski et al., 2000), hydrophilic (Jésior, 2000), electrostatic (Åqvist, 1999) and sidechain (Galitzkaya et al., 2000) interactions. In addition to its important implications in disease-related research (Bellotti et al., 1998; Ironside, 1998; Brown et al., 1999, 2000; Gursky, 1999; Kienzl et al., 1999; Gursky and Alekhov, 2000), protein folding (Nöltting, 1999b; Nöltting and Andert, 2000) has been closely associated with numerous other areas, such as RNA folding energy landscapes (Chen and Dill, 2000), chemical degradation (Maleknia and Downard, 2001), cell surface sialylation (Effertz et al., 1999), design of sequences with good folding properties (Irbäck et al., 1999), evolution of structure formation (D’Alessio, 1999), protein secretion (Chambert and Petit-Glatron, 1999), high-level protein expression (Hardesty et al., 1999), chaperone-assisted folding (Kawata et al., 1999; Gutsche et al., 2000) and macromolecular crowding inside the cell (Ellis and Hartl, 1999; van den Berg et al., 2000).

It has been demonstrated convincingly (Goto and Aimoto, 1991; Fersht et al., 1992; Dill et al., 1993; Karpus and Weaver, 1994; Orense et al., 1994; Abkevich et al., 1995; Govindarajan and Goldstein, 1995; Hamada et al., 1995; Itzhaki et al., 1995; Nöltting et al., 1995; Fersht, 1995a, b; Gross, 1996; Kuwajima et al., 1996; Unger and Moult, 1996; Wolynes et al., 1996; Gruebele, 1999; Forge et al., 2000; Grikio, 2000; Niggemann and Steipe, 2000) that the protein folding rate is directly related to the three-dimensional structure. Topology packing (Efimov, 1999; Grigoriev et al., 1999; Clementi et al., 2000) has emerged as an important concept in understanding the relationship between protein structure and its folding process. One of the key questions in the studies of structural contacts is the interplay between short- and long-range interactions during the folding reaction (Tanaka and Scheraga, 1975, 1977; Gromiha and Selvaraj, 1997, 1999; Goto et al., 1999). It has been shown that the folding rate constant, $k_f$, of a protein greatly depends on the contact order, which is a measure of complexity of the protein molecule.
chain topology (Doyle et al., 1997; Chan, 1998; Jackson, 1998; Plaxco et al., 1998; Alm and Baker, 1999; Baker and DeGrado, 1999; Muñoz and Eaton, 1999; Riddle et al., 1999; Baker, 2000; Grantcharova et al., 2000). Proteins with a simple native chain topology, i.e. of which the native structure is dominated by contacts of residues near in sequence (Fig. 1a), are predicted to fold much faster than those with a complicated chain topology (Fig. 1b), i.e. of which the native structure contains many contacts of residues remote in sequence.

To advance our understandings of the mechanism and extreme rate of protein folding as well as to provide a basic yet applicable strategy in the rational design of novel artificial proteins and re-engineering of those slow-folded proteins that often involve aggregating intermediates, it is important to address the question of how structural contacts build up during folding at a sub-nanometer resolution (Nölting, 1998, 1999a). We showed here that $-\log k_f$ correlates well with the so-called chain topology parameter, $CTP$. The definition of $CTP$ is similar to the contact order used previously to describe the complexity of the chain topology of the protein molecule, but yields much improved results. Without the need of splitting the data into subsets, e.g. $\alpha$-helix or sheet, the relationship of $-\log k_f \sim CTP$ is valid over a range of rate constants of $10^{-1} \text{s}^{-1} \leq k_f \leq 10^8 \text{s}^{-1}$ with a correlation coefficient of up to $\approx 0.87$. Using this finding, very fast rate constants may be predicted several orders of magnitude better than with previous methods.

Fig. 1. (a) Inter-residue contact map of a rapid-folding protein, $\lambda$-repressor. (b) Inter-residue contact map of a slow-folding protein, acylphosphatase. (c) Inter-residue contact map of the prion protein domain PrP (121–231).
2. Materials and methods

2.1. Calculating measures of chain topology

Inter-residue contacts were calculated at a cut-off distance of 4 Å unless stated otherwise. With the exception of some data in Fig. 3, no contacts of hydrogen atoms were included in the calculations. The following 22 protein coordinates were taken from the Brookhaven National Laboratory Protein Data Bank (Abola et al., 1997): bovine acyl-coenzyme A binding protein (2ABD), mouse prion protein domain PrP(121–231) (1AG2) (only used in Fig. 1c; available binding protein (2ABD), mouse prion protein domain 2.2. Folding rate constants

Eighteen rate constants (kf) were obtained from reference (Jackson, 1998). The kf of the 16-residue β-hairpin was chosen as previously described in Muñoz and Eaton (1999). The kf of the 10-residue helical polyalanine peptide was estimated using data in Williams et al. (1996), Gruebele (1999), Zhou and Karplus (1999), Nöltling (1999b). The kf values for the thermostable variant of λ-repressor and for the Engrailed Homeodomain, ≈50,000 s⁻¹, and 37,000 s⁻¹ are from Burton et al. (1996, 1997), and Mayor et al. (2000), respectively. kf is in units of s⁻¹ throughout this article.

3. Results and discussion

Protein folding is a surprisingly fast and extremely efficient process (Fersht et al., 1992; Fersht, 1995a, b, 1998, 2000; Nöltling et al., 1995; Nöltling, 1999b; Nöltling and Andert, 2000; Benitez-Cardoza et al., 2001; Bu et al., 2001; Ferguson et al., 2001; Galzitskaya et al., 2001; Navea et al., 2001; Ozkan et al., 2001; Rami and Udgaonkar, 2001; Roumestand et al., 2001; Volk, 2001; Vu et al., 2001; Galzitskaya et al., 2002; Kaushik et al., 2002; Yang et al., 2002).

3.1. Correlation of −log kf with CTP

It was found that −log kf correlates well with the so-called chain topology parameter, CTP:

\[ −\log k_f \sim CTP, \quad CTP = \frac{1}{LN} \sum \Delta S_{ij}^2 \]  

where kf is the rate constant of folding, L is the number of residues of the protein (chain length), N is the number of inter-residue contacts in the protein molecule, ∆Sij is the separation in sequence between the contacting residues number i and j, and “~” indicates a linear correlation.

The reasoning for the Eq. (1) started with the observation of a curvature of CO versus −log kf. It was tested which power of ∆Sij fits the best the rates −log kf. The result was close to 2 and then rounded to 2 for reasons of simplicity. Previous studies could not motivate any other exponent than 1. This may be because of smaller data sets and especially because of the absence of some peptides in the data set. The conclusion from the observation of this relation (Eq. (1)) is that the distance separation of contacts contributes nonlinearly to the free energy difference between unfolded and transition state, \( \Delta G_{U-TS} = RT \ln(hk_f/(k_B T)) \), where \( k_B = 1.3807 \times 10^{-23} \text{ J K}^{-1} \) is the Boltzmann constant, \( h = 6.6261 \times 10^{-34} \text{ J s} \) is the Planck constant, T is the absolute temperature, and \( R = 8.3145 \text{ J mol}^{-1} \text{ K}^{-1} \) is the molar gas constant (see e.g. Fersht, 1998; Nöltling, 1999b).

For a set of the 20 proteins plus a 16-residue β-hairpin (Muñoz and Eaton, 1999) and a 10-residue helical polyalanine peptide, the correlation coefficient, R, is 0.86 at a cut-off distance of 4 Å for the calculation of the inter-residue contacts of (Fig. 2a). Within the range of 10⁻⁴ s⁻¹ ≤ kf ≤ 10⁸ s⁻¹, the prediction of kf is up to a million times better than the prediction of kf using previous methods, e.g. the contact order, CO (Fig. 2b; Baker, 2000).

Fig. 2b shows that the contact order, CO, works particularly poorly for the two peptides. However, for the smaller data set of previous studies without peptides, the difference between the correlation coefficients for CO and CTP is much smaller. There have been recent
reports of similar empirical topology metrics related to folding rate (see Refs. in Makarov and Plaxco, 2003), but none of them work for isolated helices and hairpins as good as CTP.

3.2. Fit stability

For various cut-off distances from 3.5 to 8 Å, the correlation coefficient, R, for $-\log k_f$ ~ CTP is 0.80–0.87 (Fig. 3). Ignoring the contacts involving hydrogen atoms of which the positions usually are less precisely known or not fixed, e.g. due to rapid molecular motions, causes only little if any effect on R (Fig. 3). When ignoring the data points for the small peptides, the R for $-\log k_f$ ~ CTP is still 0.75–0.81 for this range of cut-off distances. The correlation thus is relatively stable regardless how the data points were selected.

There is, in general, an extensive literature suggesting a larger cut-off distance, $R_c$. Jernigan and Bahar (1996), for example, have shown that the Lennard–Johes potential between $\beta$-carbons decays to zero at around $R_c = 7.5$ Å. If we add another approximately 1 Å to account for $C_\beta$–$C_\alpha$ distance, the reasonable $R_c$ between $\alpha$-carbons seems to be 8.5 Å. Fig. 2c shows the fit for this cut-off distance. The correlation coefficient is still 0.86. Also here the correlation is not as good for the smaller data sets without peptides of previous studies. In fact, for the smaller previous data sets without peptides, the difference between the correlation coefficients for CO and CTP may not be significant.

3.3. Predictions of other folding rate constants

For a simple helical structure with very few long-range contacts, CTP ≈ 2, and the $k_f$ is predicted to be $10^6$–$10^7$ s$^{-1}$, which agrees well with the estimated diffusion limit of folding of 1 μs$^{-1}$ (Hagen et al., 1996) as well as the early groundbreaking measurements of helix–coil transitions (Hammes and Roberts, 1969).

3.4. Other factors affecting the folding rate constants

Clearly, there must be other factors affecting the rate of folding. For example, members of the fatty acid binding protein family have been shown to have folding rates that differ by three orders of magnitude, and yet these proteins have backbone structures that superimpose with RMSD values of 1 Å or less (Burns et al.,
Intuitively, these proteins would have similar CTP values and similar predicted folding rates; clearly, this is not the case. Some single or double mutations that cause little if any change in CTP can cause significant changes in folding rate constants (Nölting et al., 1997). This suggests that other strong interactions, e.g. charge interactions, are a further important structural determinant of the folding rate. $-\log k_f$ correlates comparably well ($R = 0.73$) with the number of the residues that belong to $\beta$-sheets. This may be explained by the larger number of long-range secondary structure contacts in sheets relative to helices. A weak correlation is also found between $-\log k_f$ and the chain length, $N$ (not shown).

The nature of correlation between CTP might be questionable due to dramatic stratification of fast folders from slow folders. It is known that $\alpha$-helical proteins fold faster, while $\beta$-proteins do slower. It could be possible that all of the measures of topology of the native states just reflect the secondary structure organization of proteins. To answer this question, the CTP values in Fig. 2c are highlighted according to the $\alpha$-helix content of the corresponding proteins and peptides. One can see that the three different groups of data points, 0–20% $\alpha$-helix content, 20–40% $\alpha$-helix content, and 40–100% $\alpha$-helix content, respectively, show a similar tendency like the whole data set. However, the number of available data points is still too small for a meaningful comparison of the correlation coefficients.

### 3.5. Topology of the transition state

It seems that where topology must be important is at the transition states, because one would expect that the transition states conformation topologies would bear information about corresponding free-energy barriers (Fersht et al., 1992; Fersht, 1995a, b, 1998, 2000; Oliveberg and Fersht, 1996; Oliveberg et al., 1998; Oliveberg, 2001). There are several publications that have attempted to do just that (Abkevich et al., 1994; Dokholyan et al., 2000; Vendruscolo et al., 2002). Vendruscolo et al., for example, used clustering coefficient to characterize the topological properties of the transition state conformations. They used a different measure of protein topology—node betweenness—to study relative importance of amino acids to folding kinetics.

An analysis of the structural consolidation in the transition states of proteins with various folding rate constants was performed as described in Nölting (1998, 1999a), Nölting and Andert (2000). Fig. 4 shows the results of a rapid folding protein, Arc repressor; a protein with a moderate folding rate constant, src SH3; a slow folding protein, FKBP12; and a very slow folding protein, acylphosphatase. The available sets of mutants were not sufficient to make conclusive predictions for all contacts; however, it is apparent that the complexity of the topology of the transition state undergoes significant changes as a function of the folding rate constant. The faster the rate of folding the simpler the topology of contact formation is in the transition state.

### 3.6. Why is there such a relationship with the native structure?

It is strange that one would expect correlation of the native protein structure with the kinetic properties of the proteins. Native structure has little information about kinetic barriers—look, for example, at various mutants that have similar structure albeit drastically different folding rates. Of course, the argument that one may provide is that nature may have selected those sequences that have specific folding rates (see e.g. discussion in Dokholyan and Shakhnovich, 2001). Another possibility is that some protein transition states are close to their native states—CI2, for example. It is a serious conceptual issue that has been underlying the whole discussion in the field, i.e. why there is such a relationship.

$\Phi$-value analysis on a number of proteins has shown that in many cases the transition states have low $\Phi$-values in many parts of the molecule and thus low structural consolidation as measured by the free energy of consolidation (see e.g. Nölting and Andert 2000; and references therein). Thus, protein transition states close to their native states as judged by free energy appear to be not always the case. However, even some similarity of the transition state and native structures should already influence the statistics even when the transition state is still far away from the native state. The currently available $\Phi$-value data sets are not sufficient to judge if this is the only reason for the relationship.

It should be noted, however, that structural changes connected with low energy changes might remain undetected by $\Phi$-value analysis. In particular, the absence of significant structural consolidation as measured by the free energy change does not exclude the possibility of a significantly correct long-range alignment at the measured positions or even in the whole molecule. The long-range structural properties are—on the other hand—the ones which matter most for CTP and as well for the previously used contact order. Thus the relationship of chain topology could be an indication that at least the long-range alignment of the transition state structure is close to the native structure. This hypothesis is also substantiated by the discovery of the validity of the nucleation-condensation model for folding of several proteins (see e.g. Fersht, 1995b, 1998, 2000; Nölting and Andert, 2000). According to this model, the initial folding nucleus has some degree of stability only in the presence of some correct tertiary structure interactions. It appears quite possible that in
many protein folding transition states the correct tertiary structure alignment involves more parts of the molecule than only the most consolidated residues, i.e. the nucleus. This could then further contribute to the observed relationship.

3.7. Implications in the mechanism of folding

It has been revealed in this work that a large fraction of interactions of residues close in sequence in the native structure (and already in the transition state) promotes a high rate of folding (Fig. 5). Conversely, if the structure is dominated by interactions of residues remote in sequence, folding is considerably slowed down.

Eq. (1) implies that the contribution of a contact to \( \log k_f \) is proportional to the average square of the separation of the corresponding residues in sequence. The formation of a contact between N- and C-terminus in the native structure usually takes several orders of magnitude longer than the formation of a contact of two residues half as far apart (Fig. 2a). The fit to Eq. (1) yields \( -\log k_f = 7.56 - 0.895CTP \) (Fig. 2a). As expected, this kind of behavior does not resemble the result predicted for a folding mechanism via a randomly
sampling of all conformations where one would obtain $-\log k_f \approx L - 9$ ($k_f$ is in units of $s^{-1}$). We propose that this dramatic effect of distant contacts on the folding rate constant is caused by the simultaneous build-up of secondary and tertiary structure, as emphasized in the nucleation-condensation model for folding (Itzhaki et al., 1995; Fersht, 1995a, b; Nölting et al., 1997; Shakhnovich, 1997; Nölting, 1999b). It is especially interesting to point out that the effect of the chain topology on $k_f$ is consistent with the finding that folding nuclei of proteins with complex chain topologies have many non-local contacts. A nucleus that also possesses some degree of correct tertiary structure alignment further accelerates folding more efficiently than a nucleus that involves only secondary structure contacts. Intriguingly, the prion protein domain, which is known to misfold, is found to have a particularly low CTP of 4.38 (Fig. 1c). For this protein, we speculate that the absence of sufficient long-range interactions in the folding reaction promotes misfolding.

4. Conclusions

(1) $-\log k_f$ decreases approximately proportionally to the average square of the sequence separation in the native conformation within the range of at least $10^{-1} \text{s}^{-1} \leq k_f \leq 10^{5} \text{s}^{-1}$. This correlation between $-\log k_f$ and CTP is much improved than using previously published methods. It predicts the rate constants, $k_f$, of many small monomeric proteins within less than 1 order of magnitude compared to their experimental values.

(2) The effect of a more complex topology of the structural consolidation in the slow-folding proteins is already present in the transition states of folding.

(3) For the design of novel artificial enzymes and the modification of slow-folding proteins with aggregating intermediates, short sequence separations may be preferred for the establishment of stable interactions.

Acknowledgements

We gratefully acknowledge many highly important contributions to this study by Professor Alan R. Fersht.

References


