Journal of Theoretical Biology 317 (2013) 224–228

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Inter-residue interaction is a determinant of protein folding kinetics

Shanran Huang, Jitao T. Huang*

College of Chemistry, Nankai University, Tianjin 300071, China

HIGHLIGHTS

- Folding rates of proteins are determined by the inter-residue interactions.
- Each interaction is the function of distances between any two residues.
- The contacts are a simplification in the folding-resistance interaction.

ARTICLE INFO

Article history:
Received 27 June 2012
Received in revised form
17 September 2012
Accepted 2 October 2012
Available online 11 October 2012

Keywords:
Folding mechanism of protein
Folding rates
Protein collapse
Topological structure
Contact order

ABSTRACT

In recent years, there have been many breakthroughs in the prediction of protein folding kinetics using empirical and theoretical methods. These predictions focus primarily on the structural parameters in concert with contacting residues. The non-covalent contacts are a simplified model of the interactions found in proteins. Here we investigate the physico-chemical origin and derive the approximate formula $\ln k_f = a + b \times \Sigma 1/d^6$, where d is the distance between different residues of the protein structure. It achieves -0.83 correlation with experimental over 57 two- and multi-state folding proteins, indicating that protein folding kinetics is determined by the interactions between all pairs of residues. The interaction is a short-range coupling that is effective only when two residues are in close proximity, consistent with the dominant role of the contacts in determining folding rates.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

One of the most challenging open problems in computational biology is the prediction of protein structures from their amino acid sequences—the so-called the problem of protein folding. Because of the tremendous complexity of protein folding, it is essentially impossible to predict the exact structure and folding trajectories of a protein. However, the basic physics behind folding is found to be much less complicated (Baker, 2000), because theoretical works reveal that the protein folding mechanism appears to be governed by some low-resolution features of native structures (Plaxco et al., 1998; Grantcharova et al., 2001; Dobson, 2003; Finkelstein and Galzitskaya, 2004).

Baker and coworkers (Plaxco et al., 1998; Baker, 2000) found that the folding rates of 12 small, two-state folding proteins were strongly correlated with a topologic factor, i.e. contact order (CO) (correlation coefficient, $r=-0.81$). CO is the average separation between the contacting residues in a sequence. The α-class proteins with low CO were shown to have faster folding rates than the β-class proteins with high CO. Munoz and Eaton's (1999) earlier work disclosed a similar phenomenon. More recently, Gromiha and Selvaraj (2001) reported an inverse relationship ($r=-0.78$) between folding rates and long-range order (LRO) for 23 two-state folding proteins. Metiu and coworkers (Makarov et al., 2002) observed that the folding rates of 24 small proteins decreased exponentially with the growth of the numbers (n_c) of native contacts ($r=-0.89$). Segal (2009) utilized a new topological descriptor to predict folding rates for 27 proteins ($r=-0.68$).

As the representative models, the folding rates satisfy two empirical relationships, $\ln k_f = a_1 - b_1 \times \mathrm{CO}$ (Plaxco et al., 1998; 2000) and $\ln k_f = a_2 + \ln n_c - b_2 \times n_c$ (Makarov et al., 2002). The former concerns the average sequence separation $|i-j|$ between two residues i and j that are in non-covalent contact. The latter concerns the number of contacts. Both assume that the folding kinetics is closely related to these non-covalent contacts; in the ordinary course of event the cutoff distance is set to 6 Å (or 8 Å). In fact, because the distance between residues is changed gradually during folding process, there is no clear boundary between the contact and non-contact (Yuan, 2005). Reliance upon a cutoff distance for defining the contact is only a coarse-grained approximation.

Folding process is considered to be a polymer collapse driven by hydrophobic interactions with the surrounding water (Galzitskaya et al., 2008a; 2008b; Ivankov et al., 2009). During the folding or intramolecular self-assembly, the collapse is caused by all residues,

* Corresponding author. Tel.: +86 13012293863; fax: +86 22 23724833.
E-mail address: jthuang@nankai.edu.cn (J.T. Huang).

0022-5193/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2012.10.003

which occur far apart in the sequence, coming together in 3D space. As protein volume (i.e. distance between residues) is decreased, there will be very large entropy loss (Bergasa-Caceres and Rabitz, 2003; Roder and Colón, 1997). Our method connects the folding kinetics to an intramolecular interaction that depends only on the distance between residues. Unlike empirical relationships about contact, however, the model has clear physical meaning.

2. The model

Proteins fold according to the first-order rate equation,

$$-\mathrm{d}[\mathrm{U}]/\mathrm{d}t = k_f \times [\mathrm{U}], \quad (1)$$

where [U] is the concentration of unfolded proteins at time t, and k_f is the effective rate constant of a folding reaction (i.e. folding rate). The logarithms of folding rates satisfy the empirical relationship,

$$\ln k_f = a - b\Omega, \quad (2)$$

where Ω is a structural factor associated with native protein. So far, some analyses in an attempt to predict ln k_f values of proteins with their structural properties, CO, n_c and LRO, were given by Plaxco et al. (1998), Makarov et al. (2002) and Gromiha and Selvaraj (2001), respectively.

Collapse often results in a protein whose volume is less than that of the initial unfolding state during folding reaction. In a simple model, residues are represented by points located at the positions of the C_α atoms when atomic details are ignored. The Euclidean distance, d_{ij}, between residues i and j in the three-dimensional (3D) space is $d_{ij} = \sqrt{(x_i - x_j)^2 + \left(y_i - y_j\right)^2 + (z_i - z_j)^2}$, where $P_i(x,y,z)$ and $P_j(x,y,z)$ denote the coordinates of C_α atom for residues i and j, respectively.

The model developed here assumes that all interactions between any two residues of the protein have influence on the protein folding. We consider the magnitude of the interaction, ω_{ij}, is a function of the distance between the two residues:

$$\omega_{ij} = 1/d_{ij}^{\nu}, \quad (3)$$

where the power ν has the same value for all proteins in our dataset, it is determined by fitting the experimental data. A protein of n residues comprises $n(n-1)/2$ possible inter-residue interactions. The total interaction, Ω, of the protein is equal to the sum of all inter-residue interactions, i.e. $\Omega = \Sigma\omega_{ij}$. According to Eq. (3), Ω is given by:

$$\Omega = \sum 1/d_{ij}^{\nu}, (|i-j| \geq 5), \quad (4)$$

It is noteworthy however, that the interactions among close neighbors can lead to observable errors. They should be eliminated from the present calculations. In our work the cutoff is five residues, that is, the interactions are neglected if the sequence separation is less than 5. The Ω values of proteins in our database are calculated using Eq. (4) and the results are listed in Table S1.

3. Comparison with experiment

The experimental data for folding rates of 66 proteins are obtained from the reports by Ivankov and Finkelstein (2004) and Gromiha et al. (2006). The data for the more recently characterized proteins are from our earlier datasets (Huang, Cheng 2007a, 2007b, 2008). Aligned sequences with the sequence identity up to 95% BLAST cutoff are considered to be homologous proteins (BLAST alignment http://blast.ncbi.nlm.nih.gov/). 1bni, 1c8c, 1coa, 1hdn, 1hz6, 1pse, 1shf, 1shg, and 3mef are deleted from our dataset because they are homologous to 1brs, 1bnz, 1cis, 1poh, 2ptl, 1psf, 1nyf, 1aey, and 1njc, respectively. Then, our dataset comprises the ln k_f values of 57 non-homologous proteins (39 two-state folding proteins, 17 multi-state folding proteins and a short peptide) (see Table S1).

Protein structures solved by X-ray diffraction and NMR are filed in the Protein Data Bank (PDB, http://www.rcsb.org/pdb) (Deshpande et al., 2005) that contains the Cartesian coordinates for all the atoms. For each protein, the relevant PDB file is opened and αcarbon (C_α) of each residue in the sequence is represented by one line in the file.

In order to test Eq. (2), we use Ω values to predict ln k_f determined experimentally for a nonhomologous set of 57 proteins. Pearson correlation coefficient, r, is a measurement of the strength and direction of a linear relationship between two variables. Statistical significance is determined for a p-value of student's t-test; $p < 0.01$ for all tests. Strong correlation is determined for an absolute value of correlation coefficient, $|r| > 0.7$; and weak correlation $|r| > 0.6$. The linear regression analysis is performed by the R (version2.9.1; http://www.r-project.org/) and the online statistics and forecasting software (version1.1.23-r3; http://www.wessa.net/slr.wasp). Absolute contact order (aCO) and relative contact order (rCO) are derived from Baker Laboratory: Calculate the Contact Order of Proteins (http://depts.washington.edu/bakerpg/contact_order/).

Knowing the ln k_f and $\Sigma 1/d_{ij}^{\nu}$ for each protein in the set, a least squares fitting procedure of the data with Eq. (4) is carried out with ν varying from 1 to 12. The result shows that the maximal correlation can be achieved at $\nu = 6$ (see Fig. 1), the best-fit linear relationship is

$$\ln k_f = 10.3(\pm 0.59) - 0.02(\pm 0.002) \times \sum 1/d_{ij}^{6}, \quad (5)$$

with a correlation coefficient, r, of -0.83; $p < 0.0001$ (Fig. 1), indicating that protein folding kinetics is determined by the interactions between all pairs of residues. Actually, the region $\nu = 5$–7 is equally good (Fig. 1). We also compute the correlation coefficients for two-state and three-state folding proteins, respectively. A strong correlation is observed between ln k_f and $\Sigma 1/d_{ij}^{6}$ for 37 two-state folding proteins ($r = -0.82$; $p < 0.0001$), and slightly weaker correlation for 19 multi-state folding proteins ($r = -0.75$; $p = 3.3 \times 10^{-4}$). Although the equation works better for two-state folding proteins with relatively simple structure, the model is more appropriate for proteins of all kinds.

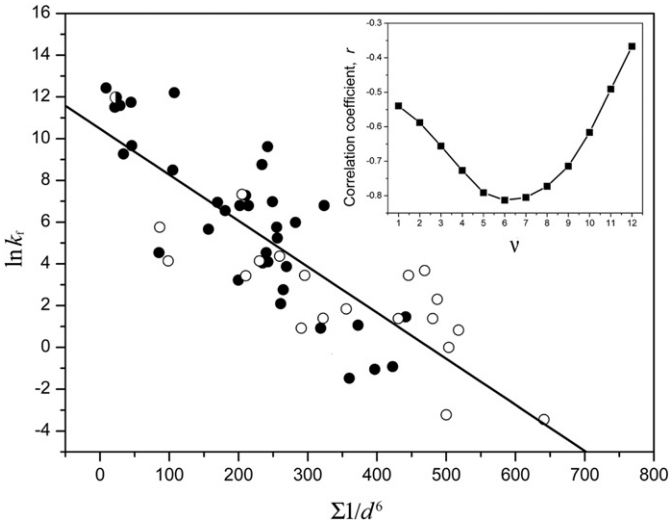

Fig. 1. Correlation between the experimentally observed folding rate constant ln k_f and the residue-residue interaction Ω, where $\Omega = \Sigma 1/d_{ij}^{\nu}$, d_{ij} is distance between residues i and j of the protein. ●, protein with two-state folding kinetics; ○, protein with multi-state folding kinetics; and ◑, short peptide. (*inset*) Correlation between ln k_f and $\Sigma 1/d_{ij}^{\nu}$ at various values of power ν. The maximal correlation is located at $\nu = 6$.

Download English Version:

https://daneshyari.com/en/article/6370935

Download Persian Version:

https://daneshyari.com/article/6370935

Daneshyari.com