Predicting protein folding rates from geometric contact and amino acid sequence

ZHENG OUYANG AND JIE LIANG
Department of Bioengineering, University of Illinois at Chicago, Chicago, Illinois 60607, USA

(RECEIVED January 22, 2008; FINAL REVISION April 4, 2008; ACCEPTED April 7, 2008)

Abstract

Protein folding speeds are known to vary over more than eight orders of magnitude. Plaxco, Simons, and Baker (see References) first showed a correlation of folding speed with the topology of the native protein. That and subsequent studies showed, if the native structure of a protein is known, its folding speed can be predicted reasonably well through a correlation with the ''localness'' of the contacts in the protein. In the present work, we develop a related measure, the geometric contact number, N_α, which is the number of nonlocal contacts that are well-packed, by a Voronoi criterion. We find, first, that in 80 proteins, the largest such database of proteins yet studied, N_α is a consistently excellent predictor of folding speeds of both two-state fast folders and more complex multistate folders. Second, we show that folding rates can also be predicted from amino acid sequences directly, without the need to know the native topology or other structural properties.

Keywords: protein folding; geometric contact number; zippers model; folding rate

In 1998, Plaxco and colleagues made the important observation that the folding rates of two-state-folding proteins correlate with the native topologies of those proteins: Faster-folding proteins tend to have more local α-helical structure, and slower-folding proteins tend to have more nonlocal β-sheet structure. Plaxco and colleagues characterized the native topology using the average relative contact order (RCO), a measure of the relative fraction of local vs. nonlocal noncovalent contacts (Plaxco et al. 1998).

Many variations of this idea have since been studied, indicating that folding rates also correlate with the long-range order (LRO) (Gromiha and Selvaraj 2001), the effective contact order (ECO) (Dill et al. 1993; Fiebig and Dill 1993), the total contact distance (TCD) (Zhou and Zhou 2002), a chain topology parameter (CTP) (Nolting et al. 2003), and the effective length of the protein, L_{eff} (Ivankov and Finkelstein 2004). A few of these quantities, such as the absolute contact order (ACO), have predictive power beyond two-state-folding rates (Ivankov et al. 2003); they also predict the rates of more complex multistate folders as well. Interestingly, although the protein's chain length was originally found to be poorly correlated with rates of two-state folders, chain length (sometimes with a fractional exponent) was later shown to correlate well with the folding rates in more recent studies (Thirumalai 1995; Finkelstein and Badretdinov 1997; Koga and Takada 2001; Galzitskaya et al. 2003; Shao et al. 2003; Naganathan and Munoz 2005).

However, these results were obtained with relatively small data sets and often begin from knowledge of the native structure of the protein (Plaxco et al. 1998; Gromiha and Selvaraj 2001; Zhou and Zhou 2002; Ivankov et al. 2003; Nolting et al. 2003). There have been several reports of predicting folding rates from protein sequences (Shao and Zeng 2003; Kuznetsov and Rackovsky 2004; Gromiha 2005; Punta and Rost 2005; Galzitskaya and Garbuzynskiy 2006), but these all require some level of information of protein structures, for example, knowledge of the structural class, or are based on prior predictions of the native secondary structures.

Reprint requests to: Jie Liang, Department of Bioengineering, University of Illinois at Chicago, 851 South Morgan Street, Room 218, Chicago, IL 60607, USA; e-mail: jliang@uic.edu; fax: (312) 996-5921.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.034660.108.

Our aim here is to develop a general method that can predict the folding rates for proteins of diverse classes based only on the amino acid sequence of the protein, without knowledge of the tertiary or secondary structures, or information of structural class, and without the aid of any other computational prediction of structural properties (e.g., secondary structures or contact order). We first use the concept of ‘‘geometric contact’’ (defined below) to study the correlation between native structure and folding rate (Li et al. 2003). Using a large set of proteins, including both two-state and multistate folders, we find that folding rates correlate well with the number of residues that form geometric contacts. The correlation coefficients are −0.86, −0.86, and −0.83 for two-state proteins, multistate proteins, and all proteins combined, respectively. Using a reduced alphabet of only two types of amino acids which are weighted differently, these correlation coefficients are all improved. The folding rate predicted from structure has a correlation coefficient of −0.86 with measured folding rate in leave-one-out jackknife tests. Based on estimated propensity values of different residues to form geometric contacts from a protein structural database, we further develop a simple algorithm that predicts folding rates from amino acid sequences alone, without any additional structural information. The predicted values correlate well with the experimental values, with a coefficient of −0.82. Our results suggest that both simple and complex proteins, over all the fold classes, may fold by a single mechanism in which spatial packing and zipping interactions are important determinants of the folding rate.

Materials and Methods

Model and data

Data set

A collected data set of experimentally determined folding rates for 80 proteins, of which 45 are two-state folders and 35 are multistate folders, were a generous gift from Ken Dill and Dr. Ke Fan (University of California at San Francisco). We have slightly modified this data set and have removed structures that contain large hetero groups, such as iron protoporphyrins, or irregular amino acids. We have also incorporated additional data from the literature. These proteins belong to different structural classes: 18 are all-α proteins, 32 are all-β proteins, and 30 are $\alpha\beta$ proteins. We took the slowest rate for multistate folders, since the faster rates are due to kinetic traps; the slowest rate corresponds to the appearance of native protein and is therefore most directly comparable with the folding rate of two-state folders. The folding rates of these proteins range over more than eight orders of magnitude, from $\ln k_f = -6.9$ for ribonucleotide isomerase (1qo2) to $\ln k_f = 12.9$ for albumin-binding domain (1prb). Tables 1 and 2 give the Protein Data Bank names and experimentally measured folding-rate values for two-state and multistate proteins, respectively. Supplemental material is available at (http://gila.bioengr.uic.edu/resources/folding/Rate.html).

Defining geometric contacts

In most studies, pairwise contacts are typically declared if two residues are within a specific cutoff distance. Such definitions can include residue pairs that have

Table 1. *The set of 45 two-state proteins*

PDB ID	Length	N_α	$\ln k_f$	Structural class
1L2Y	20	2	12.40	α
1VII	36	7	11.51	α
2PDD	43	10	9.69	α
1PRB	53	10	12.90	α
1BA5	53	4	5.91	α
1IDY	54	2	8.73	α
1FEX	59	11	8.19	α
1BDD	60	8	11.69	α
2A3D	73	7	12.7	α
1IMQ	86	39	7.28	α
1LMB	87	29	8.50	α
1ENH	54	4	10.53	α
1PGB	16	9	12.0	β
1PIN	32	19	9.37	β
1E0M	37	22	8.85	β
1E0L	37	19	10.37	β
1K9Q	40	22	8.37	β
1FMK	57	43	4.05	β
1SHG	57	45	2.10	β
1NYF	58	43	4.54	β
1PKS	76	51	−1.06	β
1C8C	64	39	6.95	β
1PSE	69	43	1.17	β
1C9O	66	47	7.20	β
1G6P	66	50	6.30	β
1CSP	67	48	6.54	β
1MJC	69	47	5.23	β
2AIT	74	61	4.21	β
1K8M	87	59	−0.71	β
1TEN	89	75	1.06	β
1FNF	90	76	−0.92	β
1WIT	93	80	0.41	β
1QTU	115	70	−0.36	β
1DIV	56	25	6.61	$\alpha\beta$
2PTL	62	36	4.10	$\alpha\beta$
2CI2	65	35	3.87	$\alpha\beta$
1RFA	78	49	7.0	$\alpha\beta$
2HQI	72	51	0.18	$\alpha\beta$
1HDN	85	51	2.69	$\alpha\beta$
1URN	96	49	5.76	$\alpha\beta$
2ACY	98	61	0.84	$\alpha\beta$
1APS	98	64	−1.47	$\alpha\beta$
1DIV	93	52	0.0	$\alpha\beta$

Table 2. *The set of 35 multistate proteins*

PDB ID	Length	N_α	$\ln k_f$	Structural class
2ABD	86	33	6.48	α
2CRO	65	26	5.35	α
1UZC	69	17	8.68	α
1CEI	85	22	5.8	α
1BRS	90	43	3.37	α
2A5E	156	85	3.50	α
1TIT	89	72	3.6	β
1FNF	93	75	5.48	β
1HNG	96	73	1.8	β
1ADW	123	84	0.64	β
1EAL	127	80	1.3	β
1IFC	131	94	3.4	β
1OPA	133	97	1.4	β
1HCD	118	78	1.1	β
1BEB	156	103	−2.20	β
1B9C	224	171	−2.76	β
1I1B	151	105	−4.01	β
1PGB	56	33	6.40	$\alpha\beta$
1UBQ	76	49	5.90	$\alpha\beta$
1GXT	89	53	4.39	$\alpha\beta$
1SCE	97	22	4.17	$\alpha\beta$
1HMK	121	57	2.79	$\alpha\beta$
3CHY	128	62	1.0	$\alpha\beta$
1HEL	129	69	1.25	$\alpha\beta$
1DK7	146	97	0.83	$\alpha\beta$
1JOO	149	75	0.30	$\alpha\beta$
2RN2	155	85	1.41	$\alpha\beta$
1RA9	159	96	−2.46	$\alpha\beta$
1PHP	175	97	2.30	$\alpha\beta$
1PHP	219	121	−3.44	$\alpha\beta$
2BLM	260	136	−1.24	$\alpha\beta$
1QOP	268	132	−2.5	$\alpha\beta$
1QOP	392	218	−6.9	$\alpha\beta$
1BTA	89	48	1.11	$\alpha\beta$
1L63	162	64	4.10	$\alpha\beta$

no steric interactions (Taylor 1997; Bienkowska et al. 1999). We take the view here that a more refined definition of geometric contact may be more useful (Li et al. 2003).

We used a contact definition based on a Voronoi criterion. Voronoi diagrams have been widely used in protein structure and folding analysis (Richards 1977; Poupon 2004). Here we illustrate our contact definition using a simple two-dimensional picture of a molecule formed by a collection of disks of uniform size (Fig. 1A). In the diagram, each Voronoi cell contains one atom, and every point inside a Voronoi cell is closer to this atom than to any other atom. A Voronoi cell is defined by its boundary edges (shown as broken lines in Fig. 1A), which are perpendicular bisectors of the line segments connecting two atom centers. For each Voronoi edge, this line segment is called the corresponding Delaunay edge (Fig. 1B). In this study, residues i and j are defined to form a geometric contact if they are connected by a Delaunay edge, and the corresponding Voronoi edge intersects with the protein body. In addition, we require that contacting residue pairs must be at least four residues apart in the primary sequence, and their spatial distance is no greater than 6.5 Å. Our parameter N_α, the geometric contact number, is simply the total number count of residues in a protein with such contacts. We first test N_α as a predictor of folding rates against other measures. The RCO was introduced by Plaxco et al. (1998):

$$RCO = \frac{1}{L \cdot N} \sum^{N} \Delta S_{i,j},$$

where N is the total number of contacts, $\Delta S_{i,j}$ is the sequence separation between residue i and j, and L is the total number of residues. RCO measures the relative importance of local and distant contacts. The ACO was also introduced by Plaxco et al. (2000):

$$ACO = \frac{1}{N} \sum^{N} \Delta S_{i,j},$$

where ACO is the average sequence separation of contacting residues, not normalized by the chain length as RCO is. Finally, chain length (L) has also been used for correlating with folding rates (Thirumalai 1995; Finkelstein and Badretdinov 1997; Koga and Takada 2001; Galzitskaya et al. 2003; Ivankov et al. 2003; Naganathan and Munoz 2005).

Results and Discussion

The Voronoi-based geometric contact definition gives an improved correlation with protein folding rates

The results of correlating folding rates $\ln k_f$ with N_α and other measures of native topology are summarized in Table 3. As others have found previously (Ivankov et al. 2003), we find that the RCO correlates poorly with folding rates for this set of 80 proteins. A better measure is the ACO. Its correlation with folding rates is $R = -0.83$ for

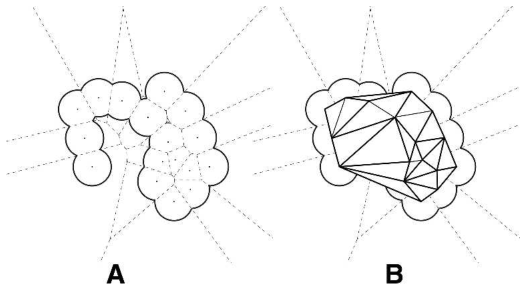

Figure 1. Voronoi diagram of a simple 2D molecule. (*A*) The molecule is formed by disks of uniform size. The dashed lines represent the Voronoi diagram, in which each region contains one atom. (*B*) The Delaunay edges of the molecule.

Table 3. *Correlation coefficients of structure-derived parameters with protein folding rates*

	Correlation coefficient		
Parameter	Two-state	Multistate	All
RCO	−0.53	0.06	−0.15
ACO	−0.83	−0.64	−0.76
L	−0.72	−0.79	−0.72
$L^{1/2}$	−0.72	−0.83	−0.78
$L^{2/3}$	−0.72	−0.82	−0.76
$L^{3/5}$	−0.72	−0.82	−0.77
Ln(L)	−0.69	−0.84	−0.79
N_α	−0.86	−0.86	−0.83

Here RCO is the contact order, ACO the absolute contact order, L is chain length, N_α is the alpha contact number.

two-state proteins, $R = -0.64$ for more complex proteins, and $R = -0.76$ for both sets combined. Previous results suggested that the protein's chain length correlates well with the folding rate (Naganathan and Munoz 2005). We found using this enlarged data set protein chain length has a strong correlation for multistate proteins ($R = -0.79$), but a weaker correlation for two-state proteins ($R = -0.72$). Although using fractional powers of the length (e.g., $L^{1/2}$, $L^{2/3}$, or $L^{3/5}$) or the logarithm ln(L) can lead to improved correlations with multistate proteins (Naganathan and Munoz 2005), they introduce little improvement for two-state proteins (see Table 3). On the other hand, the quantity N_α introduced here, correlates well in all cases ($R = -0.86$ for two-state proteins, $R = -0.86$ for multistate, and $R = -0.83$ for all 80 proteins). Figure 2 shows how these various measures correlate with folding rates of the combined set of proteins. These data indicate that an accurate description of geometric contacts improves the correlation of native protein structures with folding rates.

Comparing our geometry-based contact definition with distance-based definitions

We compare our measure using the geometric definition of contact with the following distance-based measure: We declare a pair of residues to be in contact if the distance between their C_α atoms is no greater than 6.5 Å. The results are shown in Table 4. The geometry-based definition gives a slightly better correlation than the distance-based definition for relative contact order and for our parameter of total contact number N_α, and gives the same correlation as the distance measure when using absolute contact order. More importantly, there are 8384 and 5234 pairwise contacts by the distance-based and geometry-based measures, respectively, hence 38% of the distance-based contacts either are unnecessary or degrade the correlation.

N_α is a better predictor of folding rate than chain length. Although chain length and N_α are highly correlated ($R = 0.91$), we find via subset testing that N_α is better than simple chain length at correlating with folding rates. We randomly selected a subset of 30 proteins from the 80 proteins, and carried out the correlation analysis on this subset. The correlation coefficients between the folding rate lnk_f and the geometric contact number N_α, between lnk_f and the chain length L, are recorded, respectively. This is repeated seven times. As can be seen in Figure 3, the chain length L is not a consistently good predictor of protein folding rates: The correlation R is better than −0.50 only for two subsets, and the best R-value is −0.67. Depending on the class of proteins, the R-value can be as little as −0.04. In contrast, N_α gives consistently good correlations: All are better than −0.58, with the best value being −0.79. These results suggest that N_α is more informative than chain length for understanding protein folding mechanisms.

Different geometric contacts contribute differently to folding rates

Here, we allowed each residue type i to have a weighted contribution $\boldsymbol{w}_i$, leading to the following model for protein folding rates:

$$\ln k_f = a + \boldsymbol{n}_\alpha \cdot \boldsymbol{w} \qquad (1)$$

where lnk_f is the folding rate of a protein, a is a constant, $\boldsymbol{n}_\alpha$ is a 20-dimensional vector recording the number

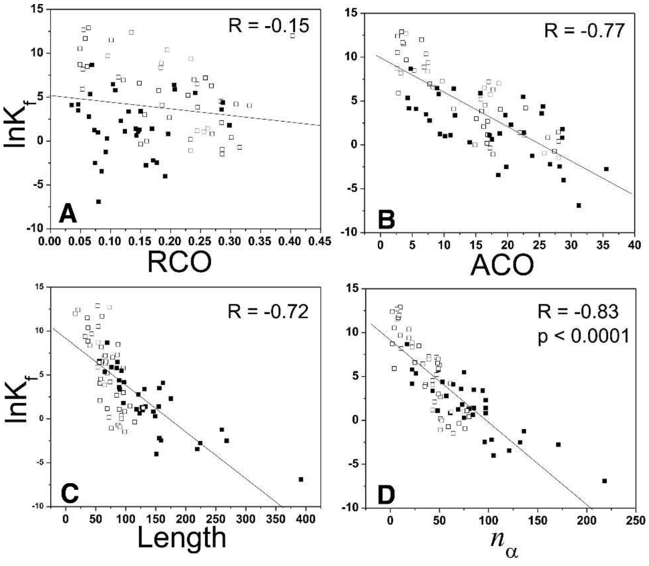

Figure 2. Relationship between different structural parameters and folding rates of two-state (open squares) and multistate (solid squares) proteins. (*A*) Relative contact order, RCO ($R = -0.15$); (*B*) absolute contact order, ACO ($R = -0.77$); (*C*) chain length ($R = -0.72$); and (*D*) N_α ($R = -0.83$).

Table 4. *Comparing distance-based and geometry-based definitions of contacts, for correlating with folding rates*

	Correlation coefficient	
Parameter	Distance contact	Geometric contact
RCO	−0.11	−0.15
ACO	−0.77	−0.77
N_α	−0.81	−0.83

The geometric definition gives a small improvement.

counts of the 20 residue types in geometric contact, and ***w*** is the 20-dimensional weight vector whose values are to be determined. Using singular value decomposition for the data set of 80 proteins, we obtain the optimal weight vector ***w***, and the baseline constant a, that minimize the residual error of the predicted lnk_f with the experimentally determined lnk_f values, by a Euclidean distance measure (Noble and Daniel 1988). The optimal weights for the 20 amino acid types are listed in Table 5. Interestingly, Val, Ile, Trp, and Tyr appear to slow down folding by the greatest extent, whereas Glu and Phe accelerate folding.

Upper bound of protein folding speed

Based on the 20 optimized weight parameters, we can estimate an upper bound for the folding speeds of the fastest proteins. In general, small proteins are fast folders: A foldable protein sequence with only 20 residues has been reported (Qiu et al. 2002). If we: (1) consider such a 20 mer, and (2) take our predicted fastest-folding residue, Glu (recognizing, however, that Glu would not lead to a stable fold), it suggests that no protein or peptide is likely to fold faster than $\ln k_f = 10.29 + 20 \times 0.451 \approx 19.3$, according to Equation 1, or roughly 4 nsec.

Folding rates and structures on a reduced alphabet of amino acids

In order to avoid overfitting, we use a reduced alphabet of amino acids containing only two types of residues, and allow these two types of residues to contribute differently to the folding rate. After exhaustive tests using different combination of residue types, we choose the following grouping of amino acids as our reduced alphabet $\mathbf{A} = (A_1, A_2)$, with $A_1 =$ (A, C, E, F, M, N, R, G, H, K, L, P, T) and $A_2 =$ (D, I, Q, S, V, W, Y). When the number counts (n_1, n_2) of residues with geometric contacts for these two reduced residue types are weighted differently with $w_1 = 0.015$ and $w_2 = -0.324$, the correlation coefficients for folding rates improves to $R = -0.87$, -0.87, and -0.87 for two state, multistate, and combined set, respectively. The resulting model $\ln k_f = 10.192 + \boldsymbol{n}_\alpha \cdot \boldsymbol{w}$ also predicts protein folding rates well. Here $\boldsymbol{n}_\alpha = (n_1, n_2)$ is the vector of number counts of geometric contact, $\boldsymbol{w} = (w_1, w_2)$ is the vector of weights. Results from jackknife tests show that predicted and measured folding rates are strongly correlated, with a correlation coefficient of 0.86 (Fig. 5A, see below).

Predicting protein folding rates from sequences

As proteins are generally tightly packed, one may assume, to first approximation, that each residue of a specific type has the same probability of a geometric contact as any other residues of the same type. With this assumption, the folding rate of a protein can be determined from knowledge of its sequence and each amino acid's general ability to form geometric contact.

The geometric contacting propensity can be estimated from known protein structures. Here, we used PDB-SELECT (2002 version), a nonredundant protein structure data set containing 1670 structures with pairwise sequence identity <25% (Boberg et al. 1992). The distribution of geometric native contacts for the 20 amino acid types and the corresponding relative values are shown in Figure 4A, and the propensity values are obtained after correction for residue composition (Fig. 4B). These propensity values collectively form the 20-dimensional contact propensity vector ***p***. We can derive the following model for correlating protein folding rates:

$$\ln k_f = a + \boldsymbol{n}_2 \cdot P\left(\boldsymbol{p} \circ w\right)$$

where $\boldsymbol{n}_2$ is the two-dimensional vector of the simple number counts of two different simplified residue types

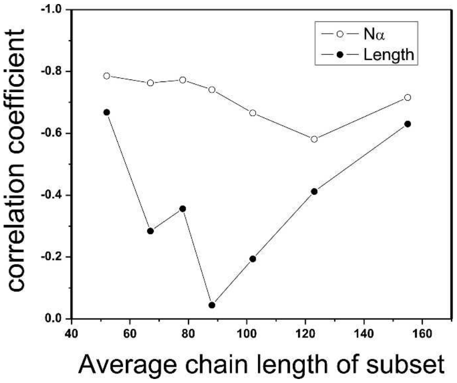

Figure 3. The geometric contact number N_α, is more robust than chain length L in correlating with protein folding rate. Results of a subset testing where 30 proteins are drawn from the original data set to form a subset. Correlation coefficients of folding rates with N_α and with the chain length L for seven such subsets are plotted.

Table 5. *The weight parameters for the different residue types in determining protein folding rates*

Residue	***w***	Residue	***w***
Glu	0.451	HIS	−0.111
Phe	0.181	THR	−0.153
Pro	0.112	GLN	−0.154
Arg	0.106	SER	−0.171
Ala	0.091	ASP	−0.179
Cys	0.069	LEU	−0.259
Gly	0.040	TYR	−0.319
Asn	0.032	TRP	−0.387
Lys	−0.043	ILE	−0.400
Met	−0.055	VAL	−0.612

w is the relative weight of each residue type, given that it is in geometric contact.

for a protein, $\boldsymbol{p}$ is the 20-dimensional geometric contacting propensity vector, $\boldsymbol{w}$ is conceptually the 20-dimensional weight vector of different contributions of the residues, "$\circ$" denotes component-wise vector product, and $P\ (\boldsymbol{p} \circ w)$ denotes the "projection" of the 20-vector of component-wide product $\boldsymbol{p} \circ w$ to the two-dimensional space of reduced alphabet, namely,

$$P(p \circ w) = \left\{ \sum\nolimits_{i \in \mathcal{A}_1} p_i \cdot w_i, \sum\nolimits_{i \in \mathcal{A}_2} p_i \cdot w_i, \right\}$$

We can denote the projection of component-wise vector product as: $\boldsymbol{w}_s = (\boldsymbol{p} \circ w)$. It integrates both the propensity of a residue type to form geometric contact and its relative contribution to folding rate. The resulting model for predicting protein folding rates is:

$$\ln k_f = 9.334 + \boldsymbol{n}_2 \cdot \boldsymbol{w}_s$$

where $\ln k_f$ is the folding rate of a protein, and $\boldsymbol{n}_2$ is the two-dimensional vector of number count of reduced residue types in the sequence of the given protein. The optimal reduced two-alphabet and values of $\boldsymbol{w}_s$ are listed in Table 6.

We find an excellent agreement between experimentally determined and predicted folding rates. The effectiveness of the model can be demonstrated in a jackknife test, in which the coefficients $\boldsymbol{w}_s$ of the model omitting one protein were calculated and the folding rate of the omitted protein is computed. The result is shown in Figure 5B, which is significantly better ($R = 0.82$) than prediction results using chain length ($R = 0.69$, Fig. 5C). As can be seen from the large amount of scattering at the right portion of Figure 5B, chain length correlates with folding rate poorly for fast folders, as folding rates of proteins of similar length (X-axis) can differ significantly. This is a phenomenon well studied in a recent theoretical work (Kachalo et al. 2006).

The deviation of sequence weights, $\boldsymbol{w}_s$, from structural weights, $\boldsymbol{w}$, can be thought of as an implicit correction by assuming some average structural information for specific residue types. Our results suggest that even models with two residue types can capture a significant amount of information about protein folding rates. This is reminiscent of the well-known HP model for studying protein stability and folding (Chan and Dill 1989; Ozkan et al. 2001; Kachalo et al. 2006).

Conclusions

We introduce here a quantity, N_α, which is a count of the number of well-packed nonlocal contacts in the native structure of a protein, where "well packed" is defined by a Voronoi criterion. The quantity N_α, is highly anticorrelated with the folding rates of 80 proteins, both two-state and multistate folders. This quantity gives a better and more consistent correlation with folding rates over this broad set of proteins than several other quantities, including the RCO, ACO, and the chain length L.

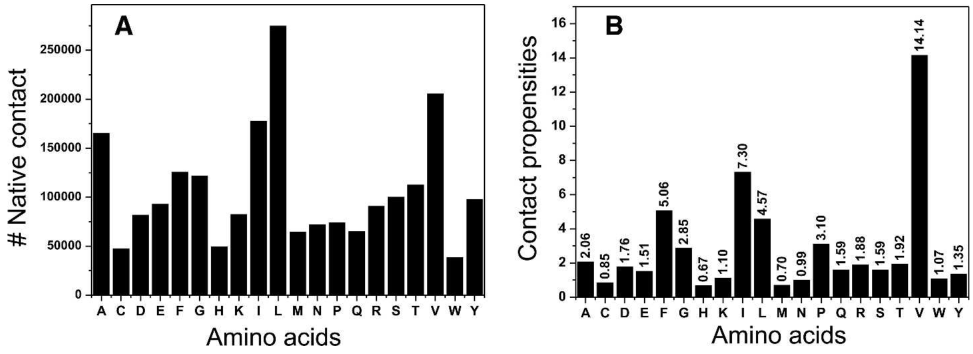

Figure 4. Propensity of residues for forming geometric contact. (*A*) Distribution of the number of native geometric contacts of 20 amino acids in the PDB select data set; (*B*) the propensity values of residues for forming geometric contact.

Table 6. *Predicting protein folding rates using reduced alphabets of amino acids*

Method	Amino acid groups	Weights	R
$\boldsymbol{w}_\alpha$	+ (A C E F M N R G H K L P T)	0.015	−0.86
	− (D I Q S V W Y)	−0.324	
$\boldsymbol{w}_s$	+ (A C E F M N R)	0.999	−0.82
	− (D I Q S V W Y G H K L P T)	−0.534	

Here, $\boldsymbol{w}_\alpha$ is the vector of weights for geometric contacts, $\boldsymbol{w}_s$ is the vector of weights for sequence, R is the correlation coefficient of leave-one-out jackknife tests; "+" denotes the group of residues which facilitate protein folding, and "−" denotes the group of residues which may retard folding.

For example, simple chain length does not correlate well with two-state folders. In addition, the correlation is not robust, as a different choice of protein samples results in a large variation in correlation (Fig. 3). The overall correlations using either RCO or ACO are not as good as that obtained from N_α. The measure N_α is not biased against shorter loops as long as their lengths are longer than a threshold of three residues, while both RCO and ACO weight more for contacts with long loops. We believe that the physical basis for this correlation is that proteins fold via a mechanism of zipping and assembly. Contacts among monomers that are more widely separated in the sequence are more difficult to form because their conformational search is more costly in chain entropy, and folding is likely to proceed through a local zipping mechanism (Dill et al. 1993; Fiebig and Dill 1993; Weikl and Dill 2003a,b; Weikl et al. 2004; Merlo et al. 2005).

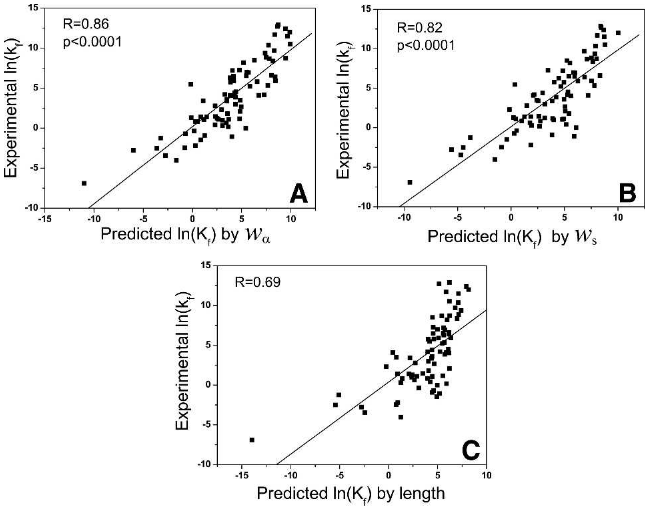

Figure 5. Scatter plots of the predicted and experimentally measured values of lnk_f in jackknife leave-one-out tests: (*A*) using weighted geometric contact number; (*B*) using sequence information only, and (*C*) using chain length.

The present work goes beyond predicting folding rates from known native structures or from known/predicted secondary structures (Ivankov and Finkelstein 2004; Gromiha 2005), and predicts rates, instead, just from the amino acid sequences of these proteins. Our prediction works even when protein sequences are based on alphabets of only two residue types. Although several previous studies can correlate protein folding rates with sequences well, they are based on a smaller data set, and they require additional structural knowledge of proteins in the form of general structural class (Kuznetsov and Rackovsky 2004; Gromiha 2005), or secondary structure information (Ivankov and Finkelstein 2004). We find that different amino acids have different propensities for folding speed. Proteins are most slowed down by Val, Ile, Trp, and Tyr forming geometric contacts, and most speeded up by Glu and Phe.

Acknowledgments

This work is supported by grants from the National Science Foundation (DBI-0646035) and the National Institutes of Health (GM079804-01A1 and GM081682 GM68958). We thank Dr. Ken Dill for stimulating discussions and for sharing the collected data of folding rates, and Dr. Martin Gruebele for helpful discussions.

References

Bienkowska, J.R., Rogers, R.G., and Smith, T.F. 1999. Filtered neighbors threading. *Proteins* **37:** 346–359.

Boberg, J., Salakoski, T., and Vihinen, M. 1992. Selection of a representative set of structures from Brookhaven Protein Data Bank. *Proteins* **14:** 265–276.

Chan, H.S. and Dill, K.A. 1989. Compact polymers. *Macromolecules* **22:** 4559–4573.

Dill, K.A., Fiebig, K.M., and Chan, H.S. 1993. Cooperativity in protein folding kinetics. *Proc. Natl. Acad. Sci.* **90:** 1942–1946.

Fiebig, K.M. and Dill, K.A. 1993. Protein core assembly processes. *J. Chem. Phys.* **98:** 3475–3487.

Finkelstein, A.V. and Badretdinov, A.Y.A. 1997. Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold. *Fold. Des.* **2:** 115–121.

Galzitskaya, O.V. and Garbuzynskiy, S.O. 2006. Entropy capacity determines protein folding. *Proteins* **63:** 144–154.

Galzitskaya, O.V., Garbuzynskiy, S.O., Ivankov, D.N., and Finkelstein, A.V. 2003. Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics. *Proteins* **51:** 162–166.

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.

Gromiha, M.M. and Selvaraj, S. 2001. Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: Application of long-range order to folding rate prediction. *J. Mol. Biol.* **310:** 27–32.

Ivankov, D.N. and Finkelstein, A.V. 2004. Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. *Proc. Natl. Acad. Sci.* **101:** 8942–8944.

Ivankov, D.N., Garbuzynskiy, S.O., Alm, E., Plaxco, K.W., Baker, D., and Finkelstein, A.V. 2003. Contact order revisited: Influence of protein size on the folding rate. *Protein Sci.* **12:** 2057–2062.

Kachalo, S., Lu, H.M., and Liang, J. 2006. Protein folding dynamics via quantification of kinematic energy landscape. *Phys. Rev. Lett.* **96:** 058106. doi: 10.1103/PhysRevLett.96.058106.

Koga, N. and Takada, S. 2001. Roles of native topology and chain-length scaling in protein folding: A simulation study with a Go-like model. *J. Mol. Biol.* **313:** 171–180.

Kuznetsov, I.B. and Rackovsky, S. 2004. Class-specific correlations between protein folding rate, structure-derived, and sequence-derived descriptors. *Proteins* **54:** 333–341.

Li, X., Hu, C., and Liang, J. 2003. Simplicial edge representation of protein structures and α contact potential with confidence measure. *Proteins* **53:** 792–805.

Merlo, C., Dill, K.A., and Weikl, T.R. 2005. ϕ Values in protein folding kinetics have energetic and structural components. *Proc. Natl. Acad. Sci.* **102:** 10171–10175.

Naganathan, A.N. and Munoz, V. 2005. Scaling of folding times with protein size. *J. Am. Chem. Soc.* **127:** 480–481.

Noble, B. and Daniel, J.W 1988. *Applied linear algebra*, 3rd ed. Prentice-Hall, New Jersey.

Nolting, B., Schalike, W., Hampel, P., Grundig, F., Gantert, S., Sips, N., Bandlow, W., and Qi, P.X. 2003. Structural determinants of the rate of protein folding. *J. Theor. Biol.* **223:** 299–307.

Ozkan, S.B., Bahar, I., and Dill, K.A. 2001. Transition states and the meaning of ϕ-values in protein folding kinetics. *Nat. Struct. Biol.* **8:** 765–769.

Plaxco, K.W., Simons, K.T., and Baker, D. 1998. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **227:** 985–994.

Plaxco, K.W., Simons, K.T., Ruczinski, I., and Baker, D. 2000. Topology, stability, sequence, and length: Defining the determinants of two-state protein folding kinetics. *Biochemistry* **39:** 11177–11183.

Poupon, A. 2004. Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. *Curr. Opin. Struct. Biol.* **14:** 233–241.

Punta, M. and Rost, B. 2005. Protein folding rates estimated from contact predictions. *J. Mol. Biol.* **348:** 507–512.

Qiu, L., Pabit, S.A., Roitberg, A.E., and Hagen, S.J. 2002. Smaller and faster: The 20-residue Trp-cage protein folds in 4 micros. *J. Am. Chem. Soc.* **124:** 12952–12953.

Richards, F.M. 1977. Areas, volumes, packing, and protein structures. *Annu. Rev. Biophys. Bioeng.* **6:** 151–176.

Shao, H. and Zeng, Z.H. 2003. A sequence function reveals new features in β-protein folding. *Protein Pept. Lett.* **10:** 435–439.

Shao, H., Peng, Y., and Zeng, Z.H. 2003. A simple parameter relating sequences with folding rates of small α-helical proteins. *Protein Pept. Lett.* **10:** 277–280.

Taylor, W.R. 1997. Multiple sequence threading: An analysis of alignment quality and stability. *J. Mol. Biol.* **269:** 902–943.

Thirumalai, D. 1995. From minimal models to real proteins: Timescales for protein folding kinetics. *J. Mol. Biol.* **5:** 1457–1469.

Weikl, T.R. and Dill, K.A. 2003a. Folding kinetics of two-state proteins: Effect of circularization, permutation, and crosslinks. *J. Mol. Biol.* **332:** 953–963.

Weikl, T.R. and Dill, K.A. 2003b. Folding rates and low-entropy-loss routes of two-state proteins. *J. Mol. Biol.* **329:** 585–598.

Weikl, T.R., Palassini, M., and Dill, K.A. 2004. Cooperativity in two-state protein folding kinetics. *Protein Sci.* **13:** 822–829.

Zhou, H. and Zhou, Y. 2002. Folding rate prediction using total contact distance. *Biophys. J.* **82:** 458–463.