Folding Rate Prediction Using Total Contact Distance

Hongyi Zhou and Yaoqi Zhou
HHMI Center for Single Molecule Biophysics, Department of Physiology & Biophysics, State University of New York at Buffalo,
Buffalo, New York 14214 USA

ABSTRACT Linear regression analysis found that either contact order (CO) or long-range order (LRO) parameter has a significant correlation with the logarithms of folding rates. This suggests that sequence separation per contact and total number of contacts are both important in determining the rate of folding. Here, the two factors are incorporated into a new parameter, total contact distance (TCD). Using a database of 28 two-state or weakly three-state folding proteins, TCD is found to be the most accurate among the three parameters (CO, LRO, and TCD) in terms of correlation and prediction. It provides even more accurate prediction than the best neural network results with two descriptors (contact order and stability per residue). The improvement is achieved in all three-structural classes (all α, β, and mixed). The accuracy of total contact distance in predicting folding rates is essentially unchanged if "short"-ranged contacts ($|i - j| \leq 14$) are not included in calculation. Thus, only long-range contacts with a sequence separation of more than 14 residues are important in determining the rate of folding. This is consistent with the results from the long-range order parameter. One of the significant outliers in prediction is found to be associated with the only protein in the database that involves nonlocal disulfide bonds. Removing the protein leads to a correlation coefficient of 0.89 between experimental observed and predicted folding rates in jackknife cross validation. The corresponding values for CO and LRO are 0.71 and 0.80, respectively.

INTRODUCTION

The logarithms of folding rates (ln k_f) of proteins that fold with two- or weakly three-state kinetics have a surprisingly simple and statistically significant correlation with a single parameter called contact order (CO) (Plaxco et al., 1998),

$$\mathrm{CO} = \frac{1}{n_c n_r} \sum_{\substack{k=1 \\ |i-j|>l_{cut}}}^{n_c} |i - j|, \tag{1}$$

where n_r is number of amino acid residues of a protein (excluding disordered regions), and n_c is number of nonlocal residue–residue contacts. A nonlocal contact is defined as two heavy atoms within a cutoff distance R_{cut} and separated by at least a residue separation cutoff value l_{cut}. Typically, $R_{cut} = 4$–6 Å and $l_{cut} = 2$ (Plaxco et al., 1998; Munoz and Eaton, 1999). The significant correlation between ln k_f and CO suggested that the average sequence separation per contact per residue of protein's native structure plays a dominant role in determining the rate of folding. This led to the development of several kinetic theories to predict folding rates from native structures (Alm and Baker, 1999; Munoz and Eaton, 1999; Debe and Goddard, 1999; Galzitskaya and Finkelstein, 1999). The accuracy in prediction can be improved further by incorporating contact order and other descriptors (such as stability) in artificial neural networks (Dinner and Karplus, 2001; Dinner et al., 2001).

Recently, a different parameter is found to correlate better with ln k_f than CO. The parameter is called long-range order (LRO) (Grombiha and Selvaraj, 2001), which is defined as the number of long-range contacts per residue. That is, LRO = n_c/n_r with $l_{cut} = 12$ and $R_{cut} = 8$ Å based on a C_α–C_α distance. The residue separation cutoff l_{cut} was also optimized for different classes of proteins ($l_{cut} = 27$, 10, and 44 for all-α, mixed α and β, and all-β proteins, respectively). The new result suggests the importance of the long-range contacts in folding kinetics. This is different from contact order in which "shorter" range contacts ($l_{cut} < 12$) also make significant contributions to the correlation with ln k_f.

In this paper, we use a new parameter, total contact distance (TCD), to predict folding rates. The new parameter is shown to be the best in correlation with the logarithms of folding rates. Its accuracy in prediction is even better than the best neural network results with two descriptors (contact order and stability per residue) (Dinner and Karplus, 2001; Dinner et al., 2001). Moreover, the accuracy in correlation is essentially unchanged for any l_{cut} values between 0 and 14. Thus, long-range contacts with a sequence separation of more than 14 residues play a dominant role in the folding rate of a protein.

THE NEW PARAMETER

The probability of finding residues i and j separated by spatial distance r, $g_{ij}(r)$, can be calculated from the equation given by (Page 55, Allen and Tildesley, 1987)

$$g_{ij}(r) = \frac{2}{n_r(n_r - 1)} \langle \delta(\mathbf{r} - \mathbf{r}_{ij}) \rangle, \tag{2}$$

where $\langle\,\rangle$ denotes ensemble average, $\delta(\mathbf{r})$ is a Dirac delta function and $n_r(n_r - 1)/2$ is the total number of pairs (a

Received for publication 30 July 2001 and in final form 11 September 2001.

Address reprint requests to Yaoqi Zhou, HHMI Center for Single Molecule Biophysics and Dept. of Physiology & Biophysics, SUNY at Buffalo, 124 Sherman Hall, Buffalo, NY 14214. Tel.: 716-829-2985; Fax: 716-829-2344; E-mail: yqzhou@buffalo.edu.

0006-3495/02/01/458/06 $2.00

normalization factor). The average sequence separation, $\bar{S}$, satisfies

$$\bar{S} = \int_0^\infty d\mathbf{r} \sum_{|i-j|>0} |i-j| g_{ij}(r). \quad (3)$$

We define TCD as the contribution to the average sequence separation by contacting residues within a distance R_{cut},

$$\text{TCD} \equiv \frac{1}{2}\int_0^{R_{cut}} d\mathbf{r} \sum_{|i-j|>0} |i-j| g_{ij}(r), \quad (4)$$

where a factor of ½ is used to simplify the expression for TCD. The term "total" is used here because it is the summation of contact distances in sequence space for all the contacts, i.e., it is not normalized by number of contacts $[\int_0^{R_{cut}} d\mathbf{r} \sum_{|i-j|>0} g_{ij}(r)]$ as in contact order.

For the case that only a single native structure is used in calculation, the integration and the ensemble average in Eqs. 2 and 4 become a simple summation over number of contacts. That is,

$$\text{TCD} = \frac{1}{n_r^2}\int_0^{R_{cut}} \sum_{|i-j|>0} |i-j| \langle \delta(\mathbf{r}-\mathbf{r}_{ij}) \rangle \, d\mathbf{r} = \frac{1}{n_r^2} \sum_{\substack{k=1 \\ |i-j|>0}}^{n_c} |i-j|, \quad (5)$$

where we have assumed $n_r \gg 1$. A more accurate evaluation of TCD would require an ensemble of native structures from molecular dynamics simulations. The summation in Eq. 5 includes both local and nonlocal contacts. To relate TCD with CO and LRO, we introduce an adjustable residue-separation cutoff, l_{cut}, in the summation. Thus, the final expression for total contact distance is

$$\text{TCD} = \frac{1}{n_r^2} \sum_{\substack{k=1 \\ |i-j|>l_{cut}}}^{n_c} |i-j|. \quad (6)$$

Compared to Eq. 1, TCD differs from CO in its prefactor. Different prefactors give different physical meanings to the two parameters. CO is a quantity per contact per residue whereas TCD is the summation over all the contacts per residue. TCD is related to CO and LRO by a simple multiplication (TCD = CO × LRO) if LRO is calculated with the same l_{cut} value as CO.

FOLDING RATE DATABASE

The database collected by Dinner et al. (2001), except the heme-containing cytochrome group and mutant proteins, is used in this study. Inclusion of these proteins in the data set, however, does not significantly alter the results reported here (see Table 2). It contains experimental data of 28 proteins. There are four all-α proteins (1LMB, 2ABD, 1IMQ, and 2PDD), 13 all-β proteins (1NYF, 1PKS, 1SHG, 1SRL, 1FNF_9, 1FNF_10, 1HNG, 1TEN, 1TIT, 1WIT, 1CSP, 1MJC, and 2AIT), and 11 mixed α,β proteins (1APS, 1HDN, 1URN, 2HQI, 1PBA, 1UBQ, 2PTL, 1FKB, 1COA, 1DIV, and 2VIK). These proteins are divided into four structurally related groups and one structurally unique group to perform structure-based cross validation as suggested by Dinner and Karplus (2001). Table 1 lists the experimental data along with CO, LRO, and TCD parameters generated from proteins' structures. As with all earlier studies, no correction for temperature variation of experimental folding rates was made.

RESULTS AND DISCUSSION

Figure 1 plots the results of linear regressions of experimental ln k_f values against CO, LRO, and TCD, respectively. Visual inspection of the figure shows that the CO has a poor correlation for proteins with high COs. A significant improvement in correlation is made by LRO, whereas the TCD is the best.

Another way to measure the significance of regression is the jackknife cross-validation method. A jackknife cross validation is done by using folding rates of all but one proteins for linear regression, and the regression parameters obtained are used to predict the folding rate of the one that was left out. The results are plotted in Fig. 2. CO makes an over estimation of folding rates for slow folding proteins and an under estimation for proteins folding with intermediate rates. The LRO parameter improves the prediction somewhat, whereas TCD provides the best results.

The results of these three parameters are quantitatively compared in Table 2. TCD has a correlation coefficient of -0.88 (p-value $= 7 \times 10^{-10}$) with the logarithms of folding rates of the 28 proteins. This is a remarkable improvement over already good correlation coefficients of -0.74 (p-value $= 7 \times 10^{-6}$) and -0.81 (p-value $= 2 \times 10^{-7}$) for CO and LRO, respectively. More importantly, the TCD parameter has the best correlation for β and mixed proteins. For example, the correlation between LRO and ln k_f (with c.c. $= -0.50$) for all-β proteins is slightly worse than that between CO and ln k_f (with c.c. $= -0.54$). The corresponding value for TCD is -0.69. For all-α proteins, the LRO is slightly better than TCD in correlation coefficients but the number of data points is too small (only 4) to be certain.

Table 2 also shows that total contact distance is the most accurate in predicting folding rates among three parameters. This is reflected from both jackknife and structure-based cross validations. A structure-based cross validation is to leave one structure group out, rather than one protein out in the jackknife test. (See Table 1 for group definition.) The prediction accuracy of total contact distance is the best in both cross-validation coefficients and root-mean-squared deviations (rmsd) between the predicted and measured ln k_f. In particular, rmsd is reduced from 2.16 for CO and 1.89 for

TABLE 1 The database and the values of three parameters (CO, LRO, and TCD) used in this study

Group*	Protein ID†	n_{r}‡	CO§ (%)	LRO¶	TCD§	Experimental ln k_{f}
All-α						
(v)	1LMB	80	18.4	0.61	0.75	8.50
(v)	2ABD	86	24.4	1.15	1.04	6.55
(v)	1IMQ	86	22.0	0.85	0.93	7.31
(v)	2PDD	43	23.9	0.49	0.75	9.80
All-β						
(i)	1NYF	58	32.9	1.40	1.22	4.54
(i)	1PKS	76	32.3	1.92	1.38	−1.05
(i)	1SHG	57	34.8	1.51	1.35	1.41
(i)	1SRL	56	34.5	1.55	1.25	4.04
(ii)	1FNF_9	90	33.9	1.99	1.30	−0.91
(ii)	1FNF_10	94	30.1	1.87	1.14	5.48
(ii)	1HNG	98	33.3	1.56	1.25	2.89
(ii)	1TEN	90	32.8	1.92	1.24	1.06
(ii)	1TIT	89	33.2	2.07	1.26	3.47
(ii)	1WIT	93	35.1	2.48	1.48	0.41
(iv)	1CSP	67	30.9	1.52	1.10	6.98
(iv)	1MJC	69	30.8	1.49	1.14	5.24
(v)	2AIT	74	34.3	2.07	1.42	4.20
Mixed α,β						
(iii)	1APS	98	34.9	2.09	1.52	−1.48
(iii)	1HDN	85	31.1	1.73	1.35	2.70
(iii)	1URN	96	30.4	1.46	1.20	5.73
(iii)	2HQI	72	31.1	2.15	1.48	0.18
(iii)	1PBA	81	29.8	1.32	1.08	6.80
(v)	1UBQ	76	29.1	1.18	1.07	7.33
(v)	2PTL	62	31.1	1.37	1.23	4.10
(v)	1FKB	107	31.6	1.98	1.30	1.46
(v)	1COA	64	31.0	1.42	1.14	3.87
(v)	1DIV	56	24.4	0.84	0.88	6.58
(v)	2VIK	126	21.7	1.67	0.97	6.80

*Groups: (i), SH3; (ii), β-sandwich; (iii), acylphosphatase; (iv), cold shock; (v), unique.

†1LMB (Burton et al., 1996), 2ABD (Kragelund et al., 1996), 1IMQ (Ferguson et al., 1999), 2PDD (Spector et al., 1998; Spector and Raleigh, 1999), 1PKS (Guijarro et al., 1988), 1FNF_9 (Plaxco et al., 1997), 1FNF_10, 1HNG, 1TIT, 1WIT (Clarke et al., 1999), 1CSP (Schindler et al., 1995; Schindler and Schmid, 1996), 2AIT (Schonbrunner et al., 1997a), 1APS (Nuland et al., 1998a), 1HDN (Nuland et al., 1998b), 1URN (Otzen et al., 1999), 2HQI (Aronsson et al., 1997), 1PBA (Villegas et al., 1995), 1UBQ (Khorasanizadeh et al., 1993), 2PTL (Scalley and Baker, 1997), 1DIV (Kuhlman et al., 1998). All others from (Jackson, 1998).

‡Only the residues that have coordinates in the PDB are counted. That is, n_{r} is number of structured residues.

§$R_{\mathrm{cut}} = 6$ Å (based on the heavy atom distance) and $l_{\mathrm{cut}} = 2$. In this table, TCD $\neq$ CO $\times$ LRO due to the use of a different cutoff for LRO.

¶The cutoffs as in the original paper (Grombiha and Selvaraj, 2001). $R_{\mathrm{cut}} = 8$ Å based on the C_{α}–C_{α} distance and $l_{\mathrm{cut}} = 12$. (Using R_{cut} based on heavy atom distance for LRO yielded essentially the same results.)

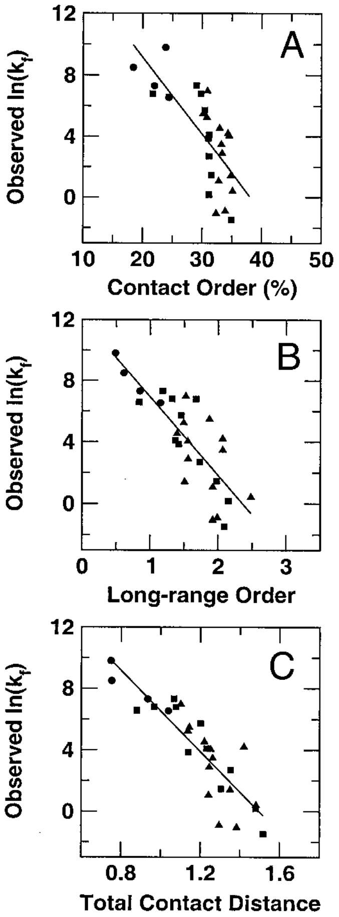

FIGURE 1 Correlation between the experimental observed ln k_{f} and the three parameters. (*a*) CO, (*b*) LRO, and (*c*) TCD. Circles denote all-α, triangles denote all-β, and squares denote mixed proteins.

LRO to 1.53 for TCD in the jackknife prediction. Similar improvement is found in structure-based cross validation.

In Table 2, we further compare the predicting power of TCD with that of genetic neural networks based on CO, stability, and other descriptors using the same data set of 33 proteins (Dinner et al., 2001). Adding cytochrome and mutated proteins does not change overall TCD results, except that the correlation with α proteins becomes worse in the presence of a few mutants. Although neural networks yielded slightly better correlations when all data are included in training, the prediction using the TCD parameter is more accurate than the two-descriptor model in structure-based cross validations and is more or at least as accurate as the three-descriptor model in terms of both correlation coefficients and rmsd values between predicted and measured ln k_{f}. This is a significant improvement considering that TCD is a quantity with only two free parameters (R_{cut} and l_{cut}).

In folding rates predicted by total contact distance (Fig. 2 *c*), one of the obvious outliers is due to the α-amylase inhibitor tendamistat (2AIT, an all-β protein). The experimental observed ln k_{f} is equal to 4.2, whereas the predicted one is only 0.70. In other words, the actual folding rate is about 33 times faster than the predicted one. A close inspection of all 28 proteins in the database found that 2AIT and another protein 2HQI (oxidized form of mercuric transport protein) are the only two proteins that have disulfide bonds. 2AIT has two nonlocal disulfide bonds at $|i - j| =$

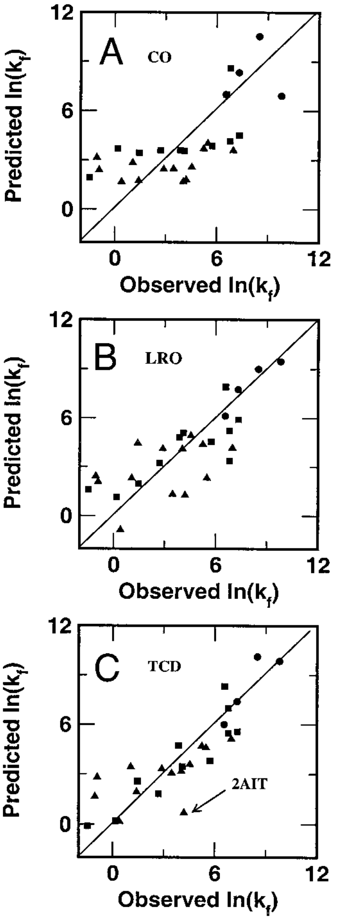

FIGURE 2 Scatter plots of the experimental observed and predicted folding rates by jackknife test. (*a*) CO, (*b*) LRO, and (*c*) TCD of contacting residues. Circles denote all-α, triangles denote all-β, and squares denote mixed proteins.

16, and 28, respectively, whereas 2HQI only has a mostly local disulfide bond ($|i - j| = 3$). Experimental studies (Schonbrunner et al., 1997b) have shown that removing one disulfide bond via mutation would reduce the folding rate of 2AIT by eight fold for the bond with $|i - j| = 16$, or 30-fold for the bond with $|i - j| = 28$. Thus, the observed folding rate after a single disulfide bond mutation ($\ln k_f = 2.1$ or 0.79) is a lot closer to the predicted one (0.70) and is within the normal prediction accuracy. However, when two disulfide bonds are removed, the protein does not fold. Thus, it seems that total contact distance cannot predict accurately the folding rates of proteins with more than one nonlocal disulfide bond.

Removing 2AIT from the database makes TCD an even better parameter in prediction. Table 2 shows that the correlation coefficient between predicted and measured $\ln k_f$ in jackknife cross validation increases from 0.86 to 0.89, and the rmsd value decreases from 1.53 to 1.40. The corresponding correlation coefficient and rmsd are 0.71 and 2.16 for CO and 0.80 and 1.84 for LRO, respectively. As for CO and LRO, significant change in prediction accuracy is not expected for the neural network three-descriptor model.

The results from neural networks (Dinner and Karplus, 2001; Dinner et al., 2001) revealed that there is a linear correlation (c.c. = 0.79) between the stability (ΔG) and $\ln k_f$ of 13 proteins that have high unnormalized contact orders ($n_r \times$ CO). They are 1PKS, 1FNF_9, 1FNF_10, 1HNG, 1TEN, 1TIT, 1WIT, 1APS, 1HDN, 1URN, 2HQI, 1FKB, and 2VIK. (Protein 2AIT was removed from the original 14 proteins for the reason described above.) The folding rates of these proteins can be described equally well or better by total contact distance (c.c. = −0.83) (c.c. = −0.75 if 2AIT is included). There is a weak correlation between ΔG and TCD (c.c. = −0.50) for these 13 proteins. It is not clear

TABLE 2 Linear regression correlation coefficients between ln k_f and three different parameters (CO, LRO, and TCD) and the results from the jackknife and structure-based cross validations. For comparison, some results from genetic neural networks (Dinner and Karplus, 2001; Dinner et al., 2001) are also included

	Correlation Coefficients				Jackknife		Structure-based	
	all	α	β	mixed	r*	δ_{rmsd}[†]	r*	δ_{rmsd}[†]
CO	−0.74 (−0.75)	−0.17	−0.52 (−0.60)	−0.71	0.70 (0.71)	2.16 (2.16)	0.68 (0.70)	3.24 (3.23)
LRO	−0.81 (−0.82)	−0.96	−0.50 (−0.56)	−0.84	0.78 (0.80)	1.89 (1.84)	0.77 (0.80)	1.95 (1.89)
TCD	−0.88 (−0.90)	−0.93	−0.69 (−0.83)	−0.92	0.86 (0.89)	1.53 (1.40)	0.84 (0.87)	1.92 (1.78)
TCD[‡]	−0.88 (−0.91)	−0.67[§]	−0.60 (−0.83)	−0.91	0.86 (0.90)	1.65 (1.43)	0.86 (0.91)	1.65 (1.41)
GNN 2[¶]	0.89[‖]				—	—	0.81	1.90**
GNN 3[††]	0.92[‖]				0.84	—	0.86	1.66**

Numbers in the parenthesis are results without 2AIT.

*Cross-validation correlation coefficient between predicted and experimental logarithmic folding rates.

[†] Rmsd between predicted and experimental logarithmic folding rates.

[‡] TCD for 33-protein data set as used in genetic neural networks.

[§] A drop in correlation is due to inclusions of cytochrome and mutated proteins.

[¶] Genetic neural networks (Dinner and Karplus, 2001; Dinner et al., 2001). Two descriptors are contact order and stability per residue.

[‖] Correlation coefficient between predicted and observed folding rates when all protein are used for training.

**Calculated from q^2_{cv} value of (Dinner et al., 2001).

[††] Three descriptors are contact order, stability per residue, and predicted β sheet contents.

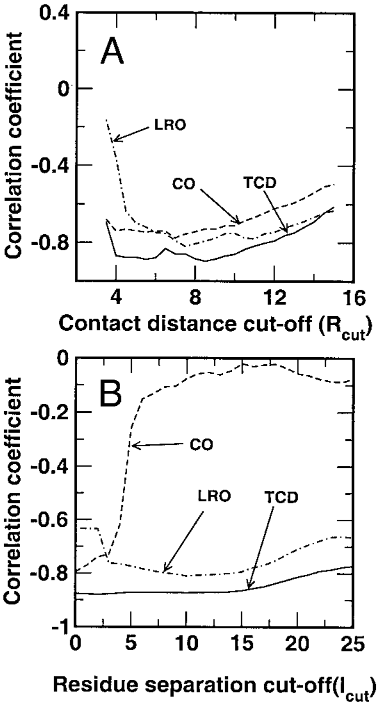

FIGURE 3 Dependence of correlation coefficient between ln k_f and CO (*dashed lines*), or ln k_f and LRO (*dashed-dotted lines*), or ln k_f and TCD (*solid lines*) on (*a*) the contact distance cutoff (R_{cut}), and (*b*) the residue separation cutoff (l_{cut}). In (*a*), $l_{cut} = 2$ for CO and TCD and 12 for LRO. In (*b*), $R_{cut} = 6$ Å for CO and TCD and 8 Å for LRO.

whether neural networks that include TCD in descriptors will continue to improve the accuracy of rate prediction.

It is of interest to know how sensitive is the result to the change of the cutoff values that are used to define a contact (R_{cut}) and the nonlocalness of the contact (l_{cut}). The dependences of correlation coefficients for CO, LRO, and TCD as a function of R_{cut} and l_{cut} are shown in Fig. 3, *a* and *b*, respectively. Results are not very sensitive to R_{cut} (up to 10 Å) for all three parameters. CO is very sensitive to the residue separation cutoff l_{cut}. The correlation is significantly worse if $l_{cut} > 4$. In contrast, correlation coefficients are stable over a wide range of l_{cut} for long-range order and total contact distance. In particular, the correlation coefficient between ln k_f and total contact distance is essentially the same for $0 \leq l_{cut} \leq 14$. Because a larger l_{cut} means that shorter-range contacts ($|i - j| \leq l_{cut}$) are not used in correlations, only long-range contacts with $|i - j| \geq 15$ determine the folding rates of proteins. This is consistent with the results from the long-range order parameter (Grombiha and Selvaraj, 2001).

To further test the performance of TCD, we apply TCD to the original database (Plaxco et al., 1998), a database of 23 proteins (Grombiha and Selvaraj, 2001), and a database of 24 proteins (Plaxco et al., 2000). TCD outperformed all other parameters in all the databases. Ladurner and Fersht (1997) and Viguera and Serrano (1997) systematically investigated the folding of CI2 and SH3 domain as a function of the length of solvent-exposed unstructured loop. The logarithms of the folding rates for those CI2 and SH3 mutants are found to have an excellent correlation with contact order (Fersht, 2000). In this case, the correlation between TCD and ln k_f is identical to that between CO and ln k_f. This is because the number of residues (excluding disordered regions) and the total number of contacts are unchanged among mutants, and total contact distance differs from contact order only by a constant factor.

What makes total contact distance to be superior to either CO or LRO in folding rate prediction? The significant correlation of either CO or LRO with ln k_f indicates that sequence distance per contact and total number of contacts per residue are both important in determining the folding rate of a protein. This is not surprising because a larger sequence distance between two residues means a greater physical distance in the coil state and a greater physical distance will take a longer time for the two residues to make contact. For a protein to fold fast, the fewer the number of rate-determining long-range contacts, the better. The new parameter, TCD, captures both effects in one parameter.

A fortran program (pdb-to-tcd.f) that calculates TCD, CO, and LRO from a pdb structure will be freely available upon request (yqzhou@buffalo.edu).

We thank Dr. M. Karplus and Dr. A. R. Dinner for providing us a preprint (Dinner et al., 2001) prior to its publication.

This work was supported by a grant from HHMI to State University of New York at Buffalo and by the Center for Computational Research and the Keck Center for Computational Biology at State University of New York at Buffalo.

REFERENCES

Allen, M. P., and D. J. Tildesley. 1987. Computer Simulation of Liquids. Oxford University Press, Oxford, U.K. 327–328.

Alm, E., and D. Baker. 1999. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl. Acad. Sci. U.S.A.* 96:11305–11310.

Aronsson, G., A.-C. Brorsson, L. Sahlman, and B.-H. Jonsson. 1997. Remarkably slow folding of a small protein. *FEBS Lett.* 411:359–364.

Burton, R. E., G. S. Huang, M. A. Daugherty, P. W. Fullbright, and T. G. Oas. 1996. Microsecond protein folding through a compact transition state. *J. Mol. Biol.* 263:311–322.

Clarke, J., E. Cota, S. B. Fowler, and S. J. Hamill. 1999. Folding studies of immunoglobulin-like β-sandwich proteins suggest that they share a common folding pathway. *Struct. Fold. Des.* 7:1145–1153.

Debe, D. A., and W. A. Goddard, III. 1999. First principles prediction of the protein folding rates. *J. Mol. Biol.* 294:619–625.

Dinner, A. R., and M. Karplus. 2001. The roles of stability and contact order in determining protein folding rates. *Nature Struct. Biol.* 8:21–22.

Dinner, A. R., S. So, and M. Karplus. 2001. Statistical analysis of protein folding kinetics. *Adv. Chem. Phys.* in press.

Ferguson, N., A. P. Capaldi, R. James, C. Kleanthous, and S. E. Radford. 1999. Rapid folding with and without populated intermediates in the homologous four-helix proteins Im7 and Im9. *J. Mol. Biol.* 286: 1597–1608.

Fersht, A. R. 2000. Transition-state structure as a unifying basis in protein-folding mechanisms: contact order, chain topology, stability, and the extended nucleus mechanism. *Proc. Natl. Acad. Sci. U.S.A.* 97: 1525–1529.

Galzitskaya, O. V., and A. V. Finkelstein. 1999. A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl. Acad. Sci. U.S.A.* 96:11299–11304.

Grombiha, M. M., and S. Selvaraj. 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.

Guijarro, J. I., C. J. Morton, K. W. Plaxco, M. Pitkeathly, I. D. Campbell, and C. M. Dobson. 1988. Folding kinetics of the SH3 domain of PI3 kinase by real-time nmr combined with optical spectroscopy. *J. Mol. Biol.* 276:657–667.

Jackson, S. E. 1998. How do small single-domain proteins fold. *Fold. Des.* 3:R81–R91.

Khorasanizadeh, S., I. D. Peters, T. R. Butt, and H. Roder. 1993. Folding and stability of a tryptophan-containing mutant of ubiquitin. *Biochemistry.* 32:7054–7063.

Kragelund, B. B., P. Hojrup, M. S. Jensen, C. K. Schjerling, E. Juul, J. Knudsen, and F. M. Poulsen. 1996. Fast and one-step folding of closely and distantly related homologous proteins of a four helix bundle family. *J. Mol. Biol.* 256:187–200.

Kuhlman, B., D. L. Luisi, P. A. Evans, and D. P. Raleigh. 1998. Global analysis of the effects of temperature and denaturant on the folding and unfolding kinetics of the N-terminal domain of the protein L9. *J. Mol. Biol.* 284:1661–1670.

Ladurner, A. G., and A. R. Fersht. 1997. Glutamine, alanine or glycine repeats inserted into the loop of a protein have minimal effects on stability and folding rates. *J. Mol. Biol.* 273:330–337.

Munoz, V., and W. A. Eaton. 1999. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci. U.S.A.* 96:11311–11316.

Nuland, N. A. J. V., F. Chiti, N. Taddei, G. Raugei, G. Ramponi, and C. M. Dobson. 1998a. Slow folding of muscle acylphosphatase in the absence of intermediates. *J. Mol. Biol.* 283:883–891.

Nuland, N. A. J. V., W. Meijberg, L. Warner, V. Forge, R. M. Scheek, G. T. Robillard, and C. M. Dobson. 1998b. Slow cooperative folding of a small globular protein HPr. *Biochemistry.* 37:622–637.

Otzen, D. E., O. Kristensen, M. Proctor, and M. Oliveberg. 1999. Structural changes in the transition state of protein folding: alternative interpretations of curved chevron plots. *Biochemistry.* 38:6499–6511.

Plaxco, K. W., K. T. Simons, and D. Baker. 1998. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* 277:985–994.

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

Plaxco, K. W., C. Spitzfaden, I. D. Campbell, and C. M. Dobson. 1997. A comparison of the folding kinetics and thermodynamics of two homologous fibronectin type III modules. *J. Mol. Biol.* 270:763–770.

Scalley, M. L., and D. Baker. 1997. Protein folding kinetics exhibit an Arrhenius temperature dependence when corrected for the temperature dependence of protein stability. *Proc. Natl. Acad. Sci. U.S.A.* 94: 10636–10640.

Schindler, T., M. A. Marahiel, and F. X. Schmid. 1995. Extremely rapid protein folding in the absence of intermediates. *Nature Struct. Biol.* 2:663–673.

Schindler, T., and F. X. Schmid. 1996. Thermodynamic properties of an extremely rapid protein folding reaction. *Biochemistry.* 35: 16833–16842.

Schonbrunner, N., K. P. Koller, and T. Kiefhaber. 1997a. Folding of the disulfide-bonded β-sheet protein tendamistat: rapid two-state folding without hydrophobic collapse. *J. Mol. Biol.* 268:526–538.

Schonbrunner, N., G. Pappenberger, M. Scharf, J. Engels, and T. Kiefhaber. 1997b. Effect of pre-formed correct tertiary interactions on rapid two-state tendamistat folding: evidence for hairpins as initiation sites for β-sheet formation. *Biochemistry.* 30:9051–9056.

Spector, S., B. Kuhlman, R. Fairman, E. Wong, J. Boice, and D. P. Raleigh. 1998. Cooperative folding of a protein mini domain: the peripheral subunit-binding domain of the pyruvate dehydrogenase multienzyme complex. *J. Mol. Biol.* 276:479–489.

Spector, S., and D. P. Raleigh. 1999. Submillisecond folding of the peripheral subunit-binding domain. *J. Mol. Biol.* 293:763–768.

Viguera, A. R., and L. Serrano. 1997. Loop length, intramolecular diffusion and protein folding. *Nature Struct. Biol.* 4:939–946.

Villegas, V., A. Azuaga, L. Catasus, D. Reverter, P. L. Mateo, F. X. Aviles, and L. Serrano. 1995. Evidence for a 2-state transition in the folding process of the activation domain of human procarboxypeptidase-A2. *Biochemistry.* 34:15105–15110.