doi:10.1016/j.jmb.2004.04.015 *J. Mol. Biol.* (2004) **339**, 951–965

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Unification of the Folding Mechanisms of Non-two-state and Two-state Proteins

Kiyoto Kamagata[1], Munehito Arai[2] and Kunihiro Kuwajima[1]*

[1]*Department of Physics, School of Science, University of Tokyo 7-3-1 Hongo, Bunkyo-ku Tokyo 113-0033, Japan*

[2]*Protein Design Research Group, Institute for Biological Resources and Functions National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba Central 6, 1-1-1 Higashi Tsukuba, Ibaraki 305-8566 Japan*

**Corresponding author*

We have collected the kinetic folding data for non-two-state and two-state globular proteins reported in the literature, and investigated the relationships between the folding kinetics and the native three-dimensional structure of these proteins. The rate constants of formation of both the intermediate and the native state of non-two-state folders were found to be significantly correlated with protein chain length and native backbone topology, which is represented by the absolute contact order and sequence-distant native pairs. The folding rate of two-state folders, which is known to be correlated with the native backbone topology, apparently does not correlate significantly with protein chain length. On the basis of a comparison of the folding rates of the non-two-state and two-state folders, it was found that they are similarly dependent on the parameters that reflect the native backbone topology. This suggests that the mechanisms behind non-two-state and two-state folding are essentially identical. The present results lead us to propose a unified mechanism of protein folding, in which folding occurs in a hierarchical manner, reflecting the hierarchy of the native three-dimensional structure, as embodied in the case of non-two-state folding with an accumulation of the intermediate. Apparently, two-state folding is merely a simplified version of hierarchical folding caused either by an alteration in the rate-limiting step of folding or by destabilization of the intermediate.

Keywords: protein folding; topology; contact order; chain length; intermediate

Introduction

Many globular proteins with more than 100 amino acid residues have an early kinetic folding intermediate with the characteristics of a molten globule state.[1–3] For some time, observation of these non-two-state folding proteins has been interpreted as evidence of a sequential mechanism of protein folding, in which the formation of the molten globule state with native-like backbone structure and compactness first takes place, and is followed by the subsequent formation of the native state with a specific tertiary structure.

However, this sequential model of protein folding has recently become a matter of serious debate. A number of small globular proteins with fewer than 100 amino acid residues have been found to exhibit two-state folding without any accumulation of an intermediate.[4] Apparently, the accumulation of a folding intermediate is not a pre-requisite for the successful folding of these proteins. A "new" view of protein folding has thus emerged, based

Abbreviations used: RCO, relative contact order; ACO, absolute contact order; ΔG^0_{UN}, the change in free energy of unfolding in the absent of a denaturant; m, a constant related to a fractional change in the solvent-accessible surface area during unfolding; Q_d, the number of sequence-distant native pairs; L, protein chain length; L_p, the total number of amino acid residues of a protein excluding the disordered terminal regions in the PDB structure; k_I, the rate constant for the formation of the folding intermediate in non-two-state proteins; r, linear correlation coefficient; ρ, population correlation coefficient; k_N, the rate constant for the formation of the native state in non-two-state proteins; r^*, partial correlation coefficient; ρ^*, population partial correlation coefficient; k_UN, the folding rate constant of two-state proteins; N, native state; U, unfolded state; I, intermediate; T, temperature; GdnHCl, guanidine hydrochloride; GuSCN, guanidine isothiocyanate.

E-mail address of the corresponding author: kuwajima@phys.s.u-tokyo.ac.jp

Table 1. Non-two-state folders

No.	Protein	L	RCO	ACO	Q_d	$\log(k_I)$	$\log(k_N)$	T	pH	Denaturant	PDB	Ref.
1	E7 colicin-binding immunity domain (Im7)	87 (85)	10.8	9.2	14	3.48	2.49	10	7.0	0 M U*	1CEI	35
2	Staphylococcal nuclease[a]	149	8.1	12.1	62	1.58	0.38	15	5.3	0 M G	1JOO	36
3	Acyl-CoA binding protein (ACBP)[b‡]	86	14.3	12.3	24.6	4.04	2.81	26	5.3	0 M G	2ABD	37
4	Intracellular lipid-binding protein (iLBP)[c]	130	13.7	17.8	78.7	1.71	0.88	–	–	–	–	38–40
5	B1 domain of protein G[‡]	57 (56)	17.3	9.7	23	3.36	2.78	20	5	0 M G*	1PGB	41
6	Interleukin-1β	153 (151)	12.4	18.7	105	0.61	−1.74	25	7.0	0 M G	1I1B	31
7	Dihydrofolate reductase (DHFR)[d]	159	14.0	22.3	107	0.70	−1.07	15	7.8	0 M U	1RA9	42
8	Barstar[e]	89	12.2	10.9	36	2.48	0.48	10	7.0	0 M U	1BTA	43
9	Apomyoglobin[f]	153 (152)	8.1	12.4	17	2.30	1.31	26	6.0	–	1DWR	44
10	Green fluorescent protein[g]	236 (227)	13.0	29.5	214	0.34	−1.20	25	7.5	–	1B9C	–
11	C-terminal first domain of the cell-surface receptor protein (CD2)	98 (97)	17.6	17.1	69.5	–	0.78	25	7	0 M G	1HNG	45
12	Phage T4 lysozyme[h]	164 (162)	7.2	11.6	28	–	1.78	25	6.0	0 M G	1L63	46
13	Ribonuclease[i]	145 (143)	13.4	18.9	83	–	0.13	25	5.5	–	–	47,48
14	Apo-pseudoazurin	123	12.5	15.4	88.5	–	0.30	15	7.0	0 M U*	1ADW	49
15	Phage 434 Cro protein	71 (65)	11.2	7.3	8	–	1.52	20	6.0	0 M U	2CRO	50
16	Barnase	110 (108)	11.4	12.3	51.7	–	1.16	25	7.5†	0 M G	1BNI	51
17	N-terminal domain of HypF	91 (88)	20.9	18.4	60.5	–	1.91	28	5.5	0 M U	1GXT	52
Proteins that exhibit only rollover in chevron plots												
18	N-terminal domain of phosophoglycerate kinase	175	11.5	20.2	105	–	1.00	25	7.5	0 M G	1PHP	53
19	Tenth fibronectin type III domain of human fibronectin (FNfn10)	94	12.2	11.5	59	–	2.39	25	5.0	0 D′ GS	1TTG	54
20	TI-I27[j]	89	17.8	15.8	66	–	1.51	25	7.4	0 M G	1TIT	55
21	Tumour suppressor p16	156	5.3	8.3	44	–	1.52	25	7.5	0 M U	2A5E	56
22	Chemotactic protein (CheY)	129 (128)	8.8	11.2	60	–	0.43	25	7.0	0 M U	3CHY	57

The columns give the following information: protein, name; L, chain length (the number in parentheses is the chain length that was used for calculating the structure-based parameters in cases for which the length differed from the full length of the protein); RCO, relative contact order (%); ACO, absolute contact order; Q_d, the number of sequence-distant native pairs; $\log(k_I)$, logarithm of the rate constant of the formation of an intermediate, k_I (s^{-1}); $\log(k_N)$, logarithm of the rate constant of the formation of the native state, k_N (s^{-1}); T, pH, and denaturant show the experimental conditions under which the folding kinetics experiments were performed (T, temperature (°C); U, urea; G, GdnHCl; GS, GuSCN; D′, molar activity; $*$, 0.4 M Na_2SO_4 was added (0.5 M Na_2SO_4 in apo-pseudoazurin); †, p^2H); PDB, Protein Data Bank code; Ref., references providing the rate constants for this study. When there were several homologous proteins for which the folding rates were available, the averaged values of the folding rate constants and the structure-based parameters of several homologous proteins were used as the rate constant values and the parameters. ‡For certain proteins, the location of the intermediate ("on" pathway or "off" pathway) in the kinetic scheme could not be determined because the rate constants of formation of the intermediate and the native states were far apart from each other, and the microscopic rate constant of formation of the intermediate, based on the on-pathway mechanism, was used (k_I). Ten proteins (1–10) exhibit multi-exponential folding kinetics in which the rate constant (k_I) for the formation of the folding intermediate has been determined experimentally, and 12 proteins (11–22) exhibit the rollover behavior in the dependence of the logarithmic rate constant of refolding on denaturant concentration although the k_I has not been determined. Among these 12 proteins, the folding intermediate has been characterized for seven (11–17) by observation of the burst phase in the kinetic refolding or by detection of the folding intermediate by pulse-labeling hydrogen exchange.

[a] The observed folding rate of the H124L mutant was used, and the PDB structure of the H124L mutant was used for the calculation of the structure-based parameters.

[b] The observed folding rate of the I86C mutant with IAEDANS was used, but the PDB structure of the wild-type protein was used for the calculation of the structure-based parameters.

[c] iLBP consisted of intestinal fatty acid binding protein (IFABP) (experimental condition: pH 7.3, 20 °C, 0.3 M GdnHCl for $\log(k_I)$ and 0 M GdnHCl for $\log(k_N)$ (PDB file: 1AEL)), ileal lipid-binding protein (ILBP) (pH 8.0, 20 °C, 0 M urea (1EAL)), and cellular retinol-binding protein II (CRBPII) (pH 8.0, 25 °C, 0 M urea (1OPA)). The $\log(k_I)$ and $\log(k_N)$ at zero denaturant concentration, except $\log(k_I)$ of IFABP, were obtained by extrapolation of the refolding limb of the chevron plots obtained from literatures to zero denaturant with a linear or a second order polynomial function of denaturant concentration. In the case of ILBP, the observed folding rate of the wild-type protein from rat was used, but the PDB structure of the wild-type porcine protein was used for the calculation of structure-based parameters.

[d] The observed folding rate of the wild-type protein was used, but the PDB structure of the N37D mutant was used for the calculation of the structure-based parameters. The $\log(k_I)$ and $\log(k_N)$ at zero denaturant concentration were obtained by extrapolation of the refolding limb of the chevron plots obtained from literatures to zero denaturant with a linear or a second-order polynomial function of denaturant concentration.

primarily on theoretical studies of protein folding,[5–8] and this new perspective has provided additional support for the idea of a simple two-state folding of globular proteins. Therefore, the sequential model of protein folding has, to a great extent, been discarded and is now often called the "classical" view of protein folding. This trend has been strengthened even further by the recent discovery that a simple empirical measure (i.e. the relative contact order (RCO)) of the native backbone topology is highly correlated with the experimental folding rate of two-state proteins.[9,10]

However, the question remains to be answered definitively: is the sequential model of protein folding truly in conflict with the simple model of two-state folding of small globular proteins? Rather surprisingly, this important question has not been addressed in great depth, most likely because it is generally assumed that non-two-state folding could be a complication caused by kinetically trapped, misfolded species that might be presumed identical with a "molten globule". Consequently, thus far, there have only been a few studies regarding the relationship between the non-two-state kinetics and native structure. These studies, including those conducted by Galzitskaya *et al.*, Micheletti, and ourselves, have shown that protein chain length,[11] absolute contact order (ACO)[12] and cliquishness[13] are all highly correlated with the observed experimental folding rate of non-two-state proteins, whereas no significant correlation of the RCO with the folding rate has been observed.[11,14] However, such studies have been concerned primarily with the formation rate of the native state of proteins. Therefore, it remains unclear whether the formation rate of the early folding intermediate of any given non-two-state protein is correlated with the native protein structure. Answering this question was therefore considered to be extremely important for understanding the role of intermediates in protein folding.

Here, we have investigated the relationships between the folding kinetics of non-two-state and two-state globular proteins and their native three-dimensional structure. To this end, we collected and analyzed the kinetic folding data for these proteins reported in the literature. As a result, we found that the rate constants for formation of folding intermediates and for the formation of the native state from an intermediate are both correlated significantly with parameters that reflect the native backbone topology. We therefore concluded that non-two-state folding, with the accumulation of a productive folding intermediate (i.e. a molten globule), is a general model for the mechanism of protein folding. Moreover, the two-state folding appears to be merely a simplified version of the more commonly observed non-two-state type of protein folding.

Results

To examine the relationship between the folding kinetics of globular proteins and their native three-dimensional structures, we collected kinetic folding data for two-state and non-two-state proteins reported in the literature. To this end, it was necessary to establish criteria for the classification of the proteins as two-state or non-two-state folders. Proteins with a heme group or disulfide bonds were excluded from the present analysis, because these may introduce additional complexity to the folding kinetics, such as heme misligations and disulfide-exchange reactions, as well as the potential restriction of conformational space caused by disulfide bonds.

Classification of proteins into two-state and non-two-state folders

The primary criterion for the classification of a given protein as either a two-state or a non-two-state folder is generally considered to be whether the folding kinetics are single-exponential or multi-exponential. However, while unfolded, certain proteins exhibit additional slow phases caused by *cis/trans* isomerizations about peptidyl-prolyl bonds in the unfolded state, and these slow phases were ignored in the analysis because they did not reflect the folding kinetics. Therefore, when the folding kinetics of a protein was multi-exponential, even after exclusion of the slow prolyl isomerizations from the kinetics, the protein was considered as a non-two-state folder. However, when a protein exhibited single-exponential kinetics, further criteria were required for determining whether the protein was a two-state or non-two-state folder.

[e] The observed folding rate of the C40A/C82A mutant was used, but the PDB structure of the wild-type protein was used for the calculation of the structure-based parameters.

[f] The PDB structure (without heme) of myoglobin from horse heart was used for the calculation of the structure-based parameters.

[g] The observed folding rate of the F99S/M153T/V163A mutant was used (S. Enoki & K. K., unpublished results), and the PDB structure of the F99S/M153T/V163A mutant was used for the calculation of the structure-based parameters. In this calculation, the chromophore that was produced by the 65th–67th residues was defined as the 66th residue.

[h] The observed folding rate of a Cys-free mutant was used, but the PDB structure of the C54T/C97A mutant was used for the calculation of the structure-based parameters.

[i] The ribonuclease consisted of ribonuclease HI from *E. coli* (0 M urea (2RN2)) and ribonuclease H from HIV-1 (0 M GdnHCl (1HRH)). In the case of ribonuclease HI, the observed folding rate of the Cys-free mutant was used, but the PDB structure of the wild-type protein was used for the calculation of the structure-based parameters.

[j] Direct tandem repeats of Ig module 27 of the I band of human cardiac titin.

When a protein switched from two-state folding to non-two-state folding, and *vice versa*, by changing the experimental conditions (e.g. temperature or pH),[4] we chose the folding mechanism (i.e. two-state or non-two-state) for which the folding behavior had already been investigated in the greatest detail.

Criteria for non-two-state folders

Single-exponential kinetics

When a protein shows single-exponential kinetics, the criterion most frequently used for judging the non-two-state behavior of protein folding has thus far been a deviation from the linear relationship of the logarithmic folding rate constant to the denaturant concentration,[11,12] and this deviation is generally referred to as "rollover". We have found 12 proteins (numbers 11–22 in Table 1) that satisfy this criterion. However, the deviation from this linearity is based on the premise that the linear relationship holds strictly for two-state proteins, and this premise may not necessarily be correct. There is no theoretical ground that leads to a strict linear relationship between the logarithmic folding rate constant and the denaturant concentration.

Therefore, we employed a more rigorous rule to identify non-two-state proteins; namely, when a protein revealed rollover behavior, we classified the protein as a non-two-state folder, but only if at least one of the following criteria was satisfied: (1) there was a burst phase (i.e. missing amplitude) in the folding kinetics; and/or (2) an early kinetic folding intermediate was present and well characterized. When the burst phase was observed in the folding kinetics, the rollover behavior was most likely caused by the presence of a folding intermediate. In general, when the folding intermediate is characterized experimentally, for example, by a pulse-labeling hydrogen-exchange method combined with NMR[15] or mass spectrometry,[16] the non-two-state folding status of the protein is considered as conclusive. Here, we found seven proteins (numbers 11–17 in Table 1) in the protein folding kinetics literature that satisfy this more rigorous rule.

Multi-exponential kinetics

When a protein shows multi-exponential kinetics, even after the exclusion of the slow steps in the unfolded state (i.e. *cis/trans* isomerizations of prolyl peptide bonds), the protein is definitely a non-two-state folder. The rate constant of the slowest phase in the multi-exponential kinetics has often been used as the rate constant of folding into the final native state, according to earlier studies.[11] However, multi-exponential kinetic behavior is derived from sequential single-pathway folding with a number of intervening folding intermediates, and from multiple parallel-pathway folding with an accumulation of intermediates along individual pathways.[17] Therefore, it remains unclear whether this type of folding is single-pathway or multiple-pathway; and it is unclear which folding phase should be considered as the folding phase to the native state when the folding type is multiple-pathway.

Here, we excluded the folding data from proteins for which the kinetic mechanism (i.e. sequential single-pathway folding or multiple parallel-pathway folding), had not been determined clearly, even in cases in which the folding kinetics were multi-exponential. As a result, we identified ten non-two-state proteins (numbers 1–10 in Table 1) in the literature, and three of which (staphylococcal nuclease, dihydrofolate reductase and green fluorescent protein) exhibited multiple parallel-pathway folding, and seven of which exhibited single-pathway folding. For the proteins with multiple parallel folding pathways, the averaged value of all folding rate constants along the multiple pathways was used as the rate constant of folding into the native state.

For all of the ten chosen multi-exponential non-two-state proteins, the rate constant of formation of the transient folding intermediate was known, and this rate constant was considered as the rate constant of folding into the intermediate. When there was more than one sequential intermediate between the initial unfolded and the final native state, we considered the folding rate constant that applied to the last intermediate.

Criteria for two-state folders

When a protein satisfied all of the following criteria, we classified the protein as a two-state folder: (1) the folding kinetics were single-exponential after the exclusion of slow steps in the unfolded state, such as the *cis/trans* isomerizations of prolyl peptide bonds; (2) there was no rollover behavior, and hence the logarithmic folding rate constant was linear as a function of the denaturant concentration; and (3) the thermodynamic parameters for the change in free energy of unfolding in the absence of a denaturant, ΔG^0_{UN}, and a constant related to a fractional change in the solvent-accessible surface area during unfolding, m, obtained from the equilibrium data, agreed with those calculated from the kinetic data.[4] However, it should be noted that U1A,[18] which showed a slight deviation from the linear relationship of the folding rate constant to the denaturant concentration due to movements in the position of the transition state, was classified into a two-state folder.

Results of the classifications

We classified 17 natural proteins (numbers 1–17 in Table 1) as the non-two-state folders; however, if we include proteins that satisfied only the condition of rollover behavior, an additional five proteins (18–22 in Table 1) could be included as well, resulting in 22 non-two-state folders (Table 1). Among these non-two-state folders, the rate constant for the formation of the transient folding intermediate has been reported for ten proteins (1–10 in Table 1). On the other hand, there were 18 proteins that were classified as two-state folders (Table 2).

Structure-based parameters of the native state

We investigated the relationships of the folding rate constants for the intermediate and native states with the structure-based parameters and a protein chain length, L, for the collected non-two-state proteins. Similarly, we investigated the relationship of the rate constant of folding into the native state using the same parameters for the two-state proteins. These structure-based parameters were as follows: (1) the absolute contact order, ACO;[19] (2) the relative contact order, RCO;[9] and (3) the number of sequence-distant native pairs, Q_d.[20] The definitions of these parameters are shown below.

The absolute contact order

The absolute contact order, ACO, is given by the arithmetical mean of the difference, $i-j$ ($i>j$), between the primary amino acid residue numbers (i and j) of two contacting residues, multiplied by n_{ij}, the number of atomic contacts between the residues, such that:

$$\mathrm{ACO} = \frac{1}{N}\sum_{i=2}^{L_p}\sum_{j=1}^{i-1}(i-j)n_{ij} \qquad (1)$$

Table 2. Two-state folders

No.	Protein	L	RCO	ACO	Q_d	$\log(k_{UN})$	T	pH	Denaturant	PDB	Ref.
1	Src homology 3 (SH3) domain[a]	69.3 (64)	18.5	11.7	39.5	1.05	–	–	–	–	58–61
2	E9 colicin-binding immunity domain (Im9)	86	12.1	10.4	19	3.16	10	7.0	0 M U	1IMQ	62
3	Cold shock protein B (CspB)[b]	66.3	17.0	11.3	49.8	2.91	25	7.0	0 M G	–	63
4	Fibronectin type III domain of human tenascin (TNfn3)	90 (89)	17.3	15.4	80	0.46	20	5.0	0 M U	1TEN	64
5	PSBD[c] †	41	11.0	4.5	2.1	4.21	41	7.9	–	2PDD	65
6	N-terminal domain of λ repressor (6–85)†	80	9.0	7.4	8.5	3.69	37	8.0	0 M U	1LMB	66
7	Immunoglobulin (IgG) binding domain of protein L[d]	62	16.3	10.1	33.7	1.78	22	7.0	0 M G	1HZ6	67
8	Activation domain of human procarboxypeptidase A2 (ADA2h)	80 (81)	14.3	11.5	33.4	2.95	25	7.0	0 M U	1O6X	68
9	Spliceosomal protein U1A[e]	102 (96)	16.8	16.1	54.3	2.50	25	6.3	0 M G	1URN	18
10	Chymotrypsin inhibitor 2 (CI2)	83 (65)	15.3	10.0	40	1.68	25	6.3	0 M G	2CI2	69
11	Histidine-containing phosphocarrier protein (HPr)	85	17.6	15.0	48	1.17	19.5	7.0	0 M G	1POH	70
12	Cold shock protein A	69	16.0	11.0	48	2.30	25	7.0	0 M U	1MJC	71
13	DNA-binding protein Sso7d[f]	64	12.7	8.1	21	3.02	20	6.1	0 M G	1C8C	72
14	Cytosolic immunity protein FKBP12	107	17.5	18.8	93	0.63	25	7.5	0 M U	1D6O	73
15	Muscle acylphosphatase (AcP)[g]	98	21.7	21.2	78.6	−0.64	28	5.5	0 M U	1APS	74
16	N-terminal domain of ribosomal protein L9	56	12.7	7.1	24	2.86	25	5.5	0 M G	1DIV	75
17	Ribosomal protein S6	101 (97)	18.9	18.4	63	2.56	25	6.3	0 M G	1RIS	32
18	Twitchin (TWIg18′)	93	20.3	18.9	100	0.18	20	5.0	0 M U	1WIT	76

The columns are almost the same as those shown in Table 1. $\log(k_{UN})$ is the logarithm of the rate constant of folding for the two-state folders, k_{UN} (s^{-1}). †The lineshape NMR analysis demonstrated that the protein folds in a two-state manner.

[a] It consisted of SH3 domains from α-spectrin (experimental condition: pH 3.5, 25 °C, 0 M urea (PDB file: 1SHG)), from chicken src (pH 6.0, 22 °C, 0 M GdnHCl (1SRL)), from phatidylinositol 3′-kinase (PI3 kinase) (pH 7.2, 20 °C, 0 M GdnHCl (1PNJ)) and from human Fyn tyrosine kinase (pH 7.2, 20 °C, 0 M GdnHCl (1SHF)). In the case of the SH3 domain from PI3 kinase, the observed folding rate of the protein with two additional N-terminal residues and four C-terminal residues was used. In the case of the SH3 domain from Fyn, the observed folding rate of the protein with two additional N-terminal residues and six C-terminal residues was used.

[b] It consisted of cold shock proteins from *Bacillus subtilis* (1CSP), from the thermophile *Bacillus caldolyticus* (1C9O), and from the hyperthermophile *Thermatonga maritima* (1G6P).

[c] Peripheral subunit-binding domain of dihydrolipoamide acetyltransferase.

[d] The observed folding rate of the Y43W mutant was used, but the PDB structure of the wild-type protein was used for the calculation of the structure-based parameters.

[e] The observed folding rate of the F56W mutant was used, but the PDB structure of the Y31H/Q36R mutant was used for the calculation of the structure-based parameters.

[f] The observed folding rate of the Y34W mutant was used, but the PDB structure of the wild-type protein was used for the calculation of the structure-based parameters.

[g] The observed folding rate of the human C21S mutant was used, but the PDB structure of the wild-type protein from horse was used for the calculation of the structure-based parameters.

where L_p is the total number of amino acid residues of a protein excluding the disordered terminal regions in the PDB structure, and:

$$N = \sum_{i=2}^{L_p} \sum_{j=1}^{i-1} n_{ij}$$

which represents the total number of contacts.[19] The two residues in contact were defined as the residues for which the inter-residue distance, measured by the distance between the two closest non-hydrogen atoms, was within 6 Å in space in the native structure.

The relative contact order

The relative contact order, RCO, is given by:[9]

$$\mathrm{RCO} = \frac{\mathrm{ACO}}{L_p} \qquad (2)$$

The number of sequence-distant native pairs

The number of native pairs, Q, is defined as:

$$Q = \sum_{i=2}^{L_p} \sum_{j=1}^{i-1} \Delta_{ij} \qquad (3)$$

where i and j are the residue numbers of two residues for which the C^α–C^α distance in space was within 6 Å in the PDB structure. $\Delta_{ij} = 1$ if $i - j > l$, and otherwise $\Delta_{ij} = 0$. The number of sequence-distant native pairs, Q_d, was defined as Q with $l = 12$.[20] The number of native pairs, $Q_{l=2}$, is referred to as $l = 2$.

Relationship between the rate constant of non-two-state folding and structure-based parameters

The rate constant for the formation of the intermediate

Figure 1 shows the relationship between the rate constant, k_I, for the formation of the folding

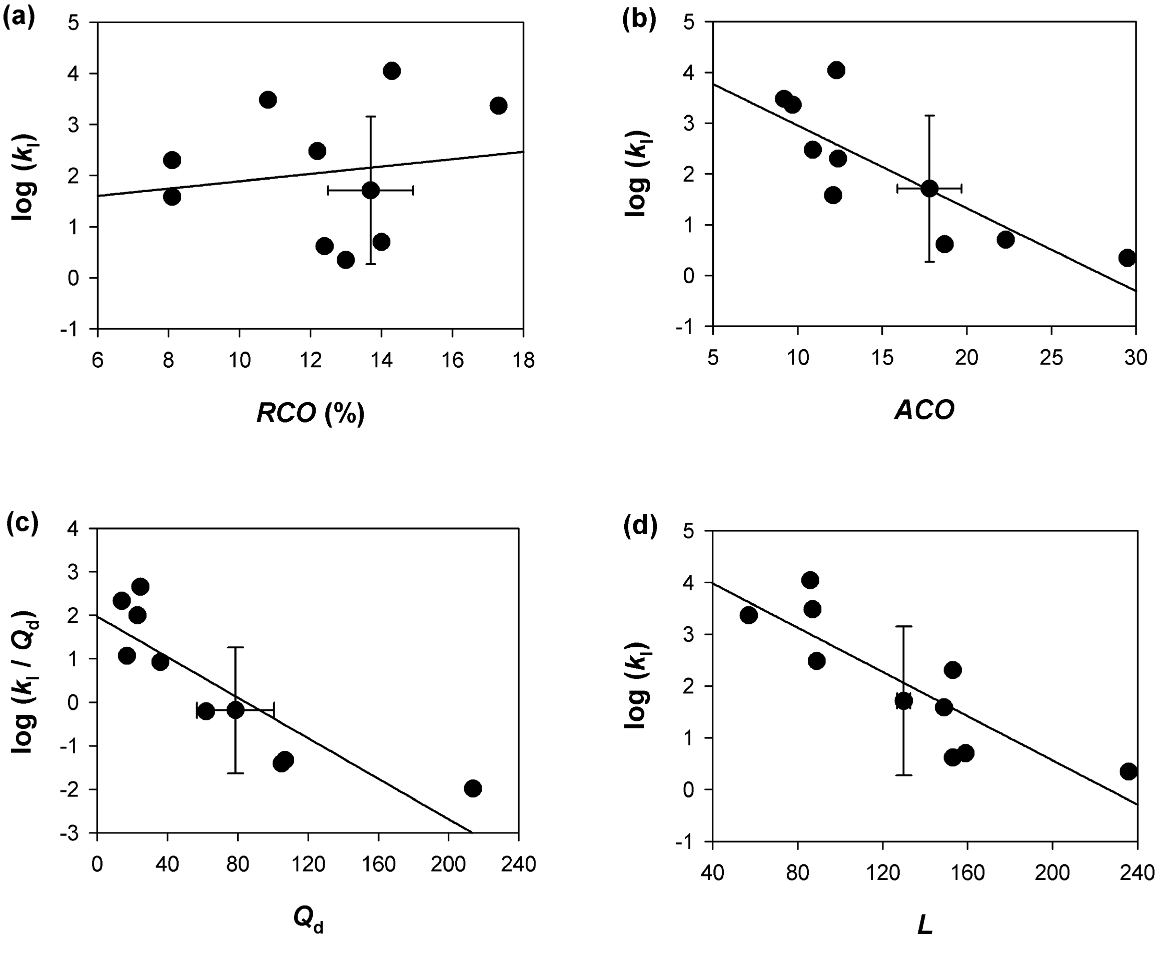

Figure 1. Logarithmic rate constant for the formation of the intermediate for non-two-state folders plotted against the following structure-based parameters: (a) relative contact order; (b) absolute contact order; (c) the number of sequence-distant native pairs; and (d) protein chain length. The continuous line represents the best linear fit for the formation of the intermediate for ten non-two-state folders. The error bar represents the scatter (standard deviation) of the values for the formation rate constant and the structure-based parameters in homologous protein, iLBP.

intermediate in non-two-state proteins, and the structure-based parameters obtained from the native structures. The logarithmic rate constant, $\log(k_I)$, was correlated strongly with the chain length, L, where the linear correlation coefficient, $r = -0.86$, and the 95% confidence interval for the population (linear) correlation coefficient estimated by Fisher's Z statistics is given by $-0.97 \leq \rho \leq -0.49$. We further investigated the relationship between $\log(k_I)$ and L^ν to find a chain-length scaling law for the formation rate of the intermediate. The observed correlation was the same when ν was between 0.1 and 1 ($r = -0.86$), whereas it was weaker when $\nu > 1$. $\log(k_I)$ correlated with $\log(L)$ ($r = -0.86$; $-0.97 \leq \rho \leq -0.50$), and its slope was -6.1 ± 1.3. The 95% confidence intervals for ρ of L or $\log(L)$ with $\log(k_I)$ does not include 0, indicating that these correlations are statistically significant. Although it was not determined which function type, L^ν ($0.1 \leq \nu \leq 1$) or $\log(L)$, gave a better prediction of $\log(k_I)$, the results indicated clearly that the formation rate of the intermediate decreased with chain length.

We found correlations between $\log(k_I)$ and the parameters that represent the native backbone topology. $\log(k_I)$ was found to be correlated strongly with ACO ($r = -0.82$; $-0.96 \leq \rho \leq -0.40$), but there was no such correlation between $\log(k_I)$ and RCO ($r = 0.16$; $-0.52 \leq \rho \leq 0.72$). $\log(k_I)$ correlated well with $Q_{l=2}$ ($r = -0.81$; $-0.95 \leq \rho \leq -0.37$, data not shown) and Q_d ($r = -0.83$; $-0.96 \leq \rho \leq -0.43$). The logarithm of k_I divided by Q_d, $\log(k_I/Q_d)$, correlated strongly with Q_d ($r = -0.87$; $-0.97 \leq \rho \leq -0.54$).[20] The 95% confidence interval for ρ of $\log(k_I)$ (or $\log(k_I/Q_d)$) with the native backbone topology (except RCO) does not include 0, indicating that these correlations are statistically significant. These results indicate that the rate of formation of the folding intermediate was determined mainly by chain length and native backbone topology.

It is possible that the significant correlation of $\log(k_I)$ with chain length has caused the correlation of $\log(k_I)$ with the native backbone topology and *vice versa*, because the parameters that represent native backbone topology are correlated well with chain length. Therefore, we calculated partial correlations between the structure-based parameters and $\log(k_I)$ (or $\log(k_I/Q_d)$), and investigated the significance of the partial correlations. The results show that the possibility that the significant correlation between $\log(k_I)$ and the native backbone topology might be due to the correlation between $\log(k_I)$ and chain length and *vice versa* cannot be ruled out (data not shown).

The rate constant for the formation of the native state

Figure 2 shows the relationship between the rate constant, k_N, for the formation of the native state from the intermediate in the case of non-two-state proteins and the structure-based parameters obtained from the native structures. The logarithmic rate constant, $\log(k_N)$, was found to correlate with L ($r = -0.71$ (-0.66); $-0.89 \leq \rho \leq -0.34$ ($-0.85 \leq \rho \leq -0.33$)), whereby the values outside and inside of the parentheses represent the linear correlation coefficients for the 17 protein and the 22 protein data set, respectively. We then investigated the relationship between $\log(k_N)$ and L^ν in order to identify a chain-length scaling law for the formation rate of the native state. The results were very similar to those observed above in the case of $\log(k_I)$. The observed correlation was the same when ν was between 0.1 and 1 ($r = -0.72$ to -0.71 (-0.67 to -0.66)), although it was weaker when $\nu > 1$. $\log(k_N)$ was correlated with $\log(L)$ ($r = -0.72$ (-0.67); $-0.89 \leq \rho \leq -0.36$ ($-0.85 \leq \rho \leq -0.34$)), and its slope was -6.1 ± 1.5 (-5.5 ± 1.4). The 95% confidence intervals for ρ of L or $\log(L)$ with $\log(k_N)$ do not include 0, indicating that these correlations are statistically significant. Although it was not determined which function type, L^ν ($0.1 \leq \nu \leq 1$) or $\log(L)$, better predicted $\log(k_N)$, the results indicated that the formation rate of the native state decreased with chain length. The present results were thus consistent with those reported previously for k_N.[11]

As observed above in the case of $\log(k_I)$, $\log(k_N)$ correlated well with ACO ($r = -0.71$ (-0.67); $-0.89 \leq \rho \leq -0.36$ ($-0.85 \leq \rho \leq -0.34$)), whereas no such correlation was observed between $\log(k_N)$ and RCO ($r = 0.11$ (0.08); $-0.39 \leq \rho \leq 0.56$ ($-0.35 \leq \rho \leq 0.49$)). $\log(k_N)$ correlated with both $Q_{l=2}$ ($r = -0.70$ (-0.67); $-0.88 \leq \rho \leq -0.32$ ($-0.85 \leq \rho \leq -0.34$), data not shown) and Q_d ($r = -0.79$ (-0.74); $-0.92 \leq \rho \leq -0.49$ ($-0.88 \leq \rho \leq -0.46$)). $\log(k_N/Q_d)$ was correlated highly with Q_d ($r = -0.85$ (-0.82); $-0.94 \leq \rho \leq -0.61$ ($-0.92 \leq \rho \leq -0.60$)).[20] The 95% confidence interval for ρ of $\log(k_N)$ (or $\log(k_N/Q_d)$) with the native backbone topology (except RCO) does not include 0, indicating that these correlations are statistically significant. Furthermore, the partial correlation between Q_d and $\log(k_N/Q_d)$ after eliminating the effect of L was statistically significant ($r^* = -0.66$ (-0.64); $-0.87 \leq \rho^* \leq -0.24$ ($-0.84 \leq \rho^* \leq -0.28$), where r^* and the interval of ρ^* represent the partial correlation coefficient and the 95% confidence interval for the population partial correlation coefficient, respectively), suggesting that the significant correlation between $\log(k_N)$ and the native backbone topology may not be due to solely the correlation between $\log(k_N)$ and chain length. On the other hand, the possibility that the significant correlation between $\log(k_N)$ and chain length might be due to the correlation between $\log(k_N)$ and the native backbone topology, cannot be eliminated (data not shown). These results were therefore very similar to those for $\log(k_I)$, and they indicated that the rate of formation from the intermediate to the native state was determined by chain length and native backbone topology.

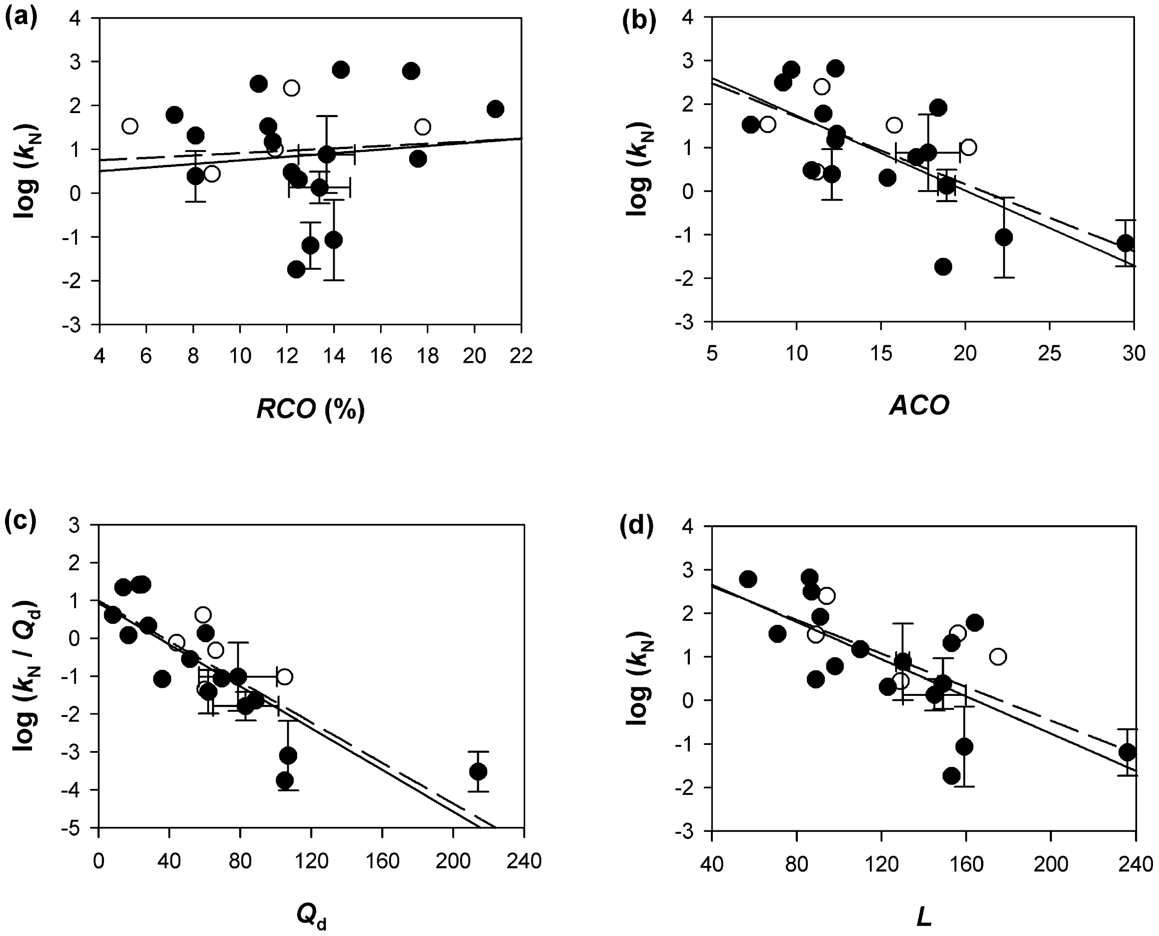

Figure 2. Logarithmic rate constant for the formation of the native state for non-two-state folders plotted against the following structure-based parameters: (a) relative contact order; (b) absolute contact order; (c) the number of sequence-distant native pairs; and (d) protein chain length. The continuous line represents the best linear fit for the formation of the native state for 17 non-two-state folders. The broken line represents the best linear fit for the formation of the native state for 22 non-two-state folders. The error bars represent the scatter (standard deviation) of the values for the formation rate constant and the structure-based parameters in homologous proteins, iLBP and ribonuclease, and in the proteins that show the folding along multiple-parallel pathways.

The relationship between the rate constant of two-state folding and the structure-based parameters

Figure 3 shows the relationships observed between the folding rate constant, k_{UN}, of two-state folders and the structure-based parameters obtained from the native structure; these relationships can be compared with those observed above for the non-two-state folding proteins. The logarithmic rate constant, log(k_{UN}), correlated only weakly with L ($r=-0.56$; $-0.82\leq\rho\leq-0.12$). log(k_{UN}) correlated well with both RCO ($r=-0.84$; $-0.94\leq\rho\leq-0.62$) and ACO ($r=-0.78$; $-0.92\leq\rho\leq-0.50$). log(k_{UN}) correlated strongly with Q_d ($r=-0.83$; $-0.94\leq\rho\leq-0.60$), although there was only a weak correlation between log(k_{UN}) and $Q_{l=2}$ ($r=-0.66$; $-0.86\leq\rho\leq-0.27$, data not shown). log(k_{UN}/Q_d) correlated strongly with Q_d ($r=-0.88$; $-0.95\leq\rho\leq-0.70$). Furthermore, the partial correlations between the native backbone topology and log(k_{UN}) (or log(k_{UN}/Q_d)) after eliminating the effect of L are statistically significant (data not shown). These results indicated that the folding rate of two-state folders is determined to a greater extent by the native backbone topology than by chain length. These observations of two-state proteins were consistent with previously reported findings.[9,10,20]

Robustness of the correlations against experimental conditions

The folding rates of the proteins under similar but slightly different experimental conditions (temperature, pH or salt concentration as listed in Tables 1 and 2) were used in this study. The slightly different conditions might affect the folding rates, and obscure the correlations between the structure-based parameters and the folding rate constants. Here, we obtained 10,000 sets of estimates of the logarithmic rate constants under the standard condition (pH 7, 25 °C, and 0 M Na_2SO_4)

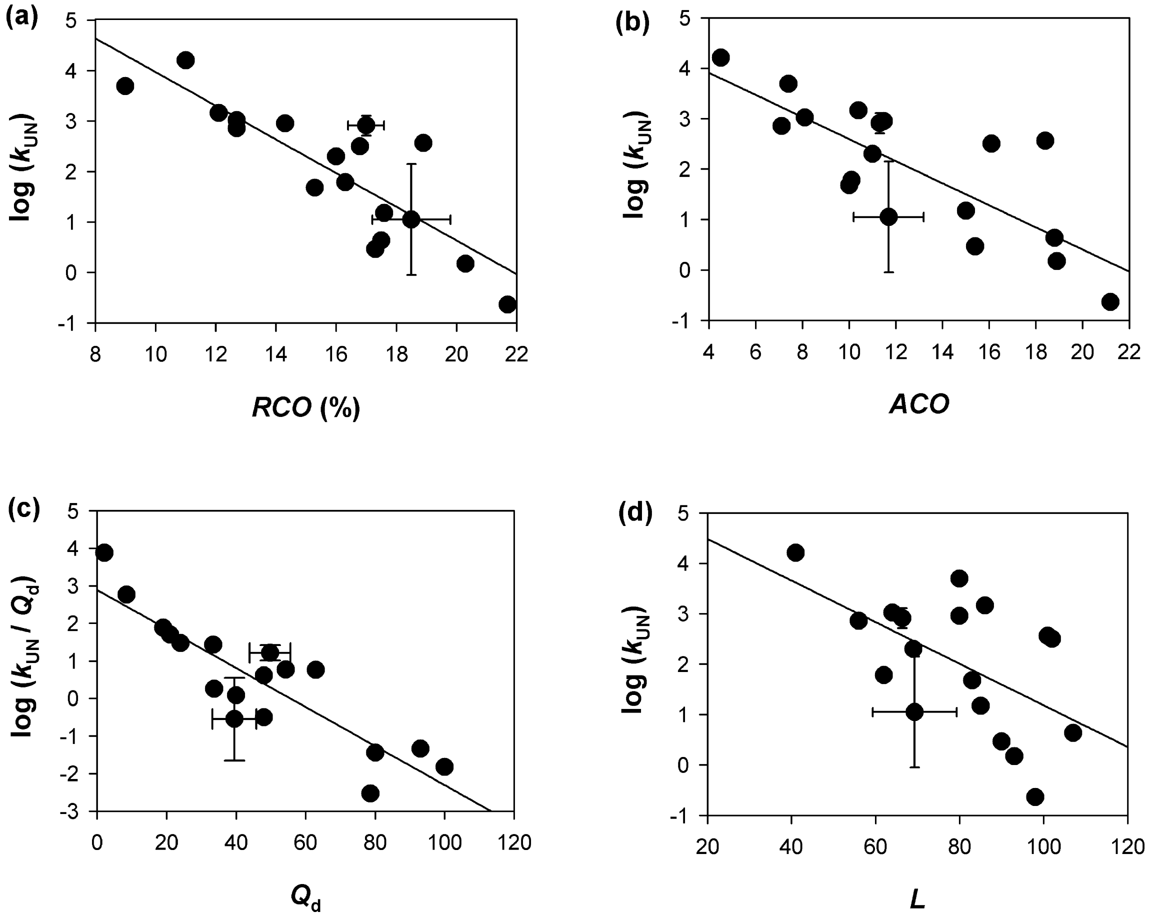

Figure 3. Logarithmic rate constant for the formation of the native state for two-state folders plotted against the following structure-based parameters: (a) relative contact order; (b) absolute contact order; (c) the number of sequence-distant native pairs; and (d) protein chain length. The continuous line represents the best linear fit for the formation of the native state for 18 two-state folders. The error bars represent the scatter (standard deviation) of the values for the formation rate constant and the structure-based parameters in homologous proteins, SH3 domains and CspB.

for the non-two-state and two-state proteins listed in Tables 1 and 2. These estimates were calculated by equation (6), which assumes that the effects of the slight deviations in pH, temperature and concentration of Na_2SO_4 on the rate constant values are random, but nevertheless proportional to these deviations (see Methods). We then investigated the correlations between these estimated folding rates and the structure-based parameters for the above 10,000 sets. As a result, we found that the slightly different conditions did not show much effect on the correlations for the non-two-state or two-state proteins, indicating the robustness of the correlations against the experimental conditions (data not shown).

Discussion

The present results demonstrated that the logarithmic rate constants, $\log k_I$ and $\log k_N$, which represent the formation of the folding intermediate (I) from an unfolded state (U) and the formation of the native state (N) from I, respectively, were both correlated significantly with protein chain length (*L*) and the structure-based parameters (ACO and Q_d); the slower the folding rate, the larger the topological parameters, ACO and Q_d. We found that another topological parameter, cliquishness, was similarly correlated with $\log k_I$ and $\log k_N$ (data not shown). Therefore, both the processes from U to I and from I to N are rate-limited by the process of forming a more native-like backbone topology. Thus, protein molecules become progressively more native-like during the folding process from U to N *via* I. The present results thus clearly demonstrate that the kinetic intermediate of refolding in the case of non-two-state folders is a real, productive folding intermediate.

We now address the possible relationship between the present results for non-two-state folders and the known results for two-state folders, and we consider the relationship of the present results to the theoretical folding predictions. Finally, we present a unified mechanism of globular protein folding.

Comparison of non-two-state and two-state proteins

Essentially, the same correlations of $\log k_N$ with ACO and Q_d as found in non-two-state proteins were observed in two-state proteins (Figures 2 and 3), but there were significant differences between non-two-state and two-state folders. The logarithmic rate constant of folding correlated significantly with RCO in the case of two-state folders, although there were no significant correlations between $\log k_N$ and RCO, nor between *log* k_I and RCO in non-two-state folders. Both $\log k_N$ and $\log k_I$ for the non-two-state proteins decreased significantly with L, whereas the decrease in $\log k_{UN}$ with L was only slightly significant in the two-state folders. Because RCO was given by ACO/L, it is extremely likely that the correlations of the logarithmic rate constants with ACO and L cancel each other out in the RCO for the non-two-state folders.

At this point, certain questions arise. For example, it remains unclear why these correlations with ACO and L did not cancel out in the case of two-state folders? Moreover, it is unclear why there were differences between the non-two-state and the two-state proteins as regards the correlations with RCO and L?

A possible answer to the above questions may be provided by the difference in the chain-length distribution between the non-two-state and two-state proteins. Figure 4 shows the chain-length distributions of the non-two-state and two-state proteins. The chain length, L, was widely distributed (i.e. from 50 to 240) in the non-two-state folders, whereas the distribution for the two-state folders was narrower (i.e. between 40 and 110). The present findings thus suggested strongly that the only slightly significant correlation between $\log k_{UN}$ and L in the case of the two-state folders may have been due to this narrow distribution of L, whereby the variation in $\log k_{UN}$ may have become comparable to the change in $\log k_{UN}$ over the range of L of the proteins.

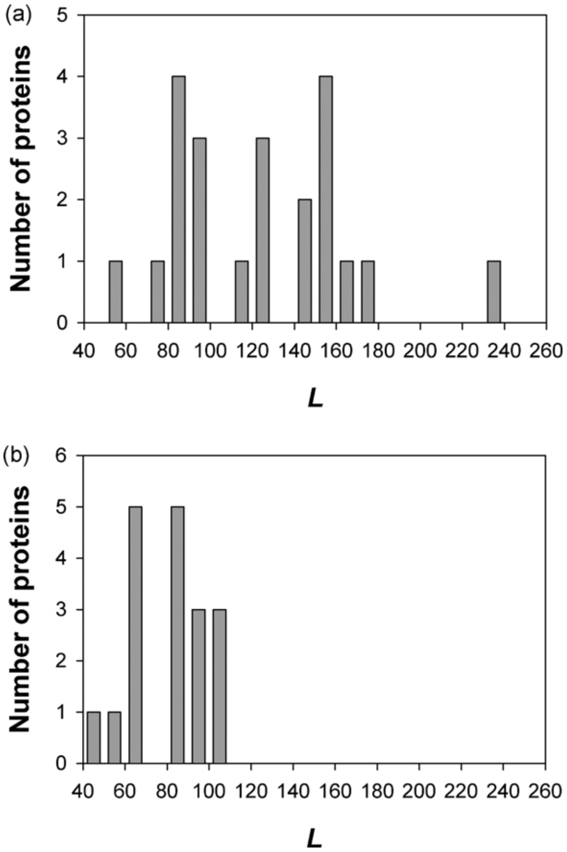

Figure 4. Distribution of protein chain length for (a) non-two-state folders and (b) two-state folders.

The *L*-dependence of the folding rate constant

The finding that both $\log k_I$ and $\log k_N$ of the non-two-state proteins decreased significantly with L has significant implications as regards our understanding of the mechanism of protein folding. In accordance with this finding, earlier theoretical studies have suggested that the folding rate becomes slower with L.[21–24] For example, Thirumalai has reported the predicted time scales for protein folding kinetics based on a multipathway mechanism using minimal models of real proteins.[21] The logarithmic rate to search a native-like state among compact unfolded states is predicted to scale as $\log L$, and the average activation barrier separating the low-energy structures and the final native state is predicted to scale as $L^{1/2}$. Based on Monte Carlo simulations of lattice proteins with various chain lengths, Gutin *et al*. demonstrated that $\log k_N$ scales as $\log L$.[22] From their thermodynamic considerations of a nucleation-and-growth mechanism of protein folding, Finkelstein & Badretdinov demonstrated that the activation barrier to form the native state scales as $L^{2/3}$.[23] On the basis of folding simulations using Go-like model proteins, Koga & Takada have reported that $\log k_N$ scales as $L^{0.6}$.[24] The present finding of the dependence of the folding rate constants, k_I and k_N, on L suggested that similar multi-pathway or nucleation-and-growth mechanisms may play a role in the folding reactions of real proteins; however, we were unable to distinguish the different mechanisms in the present analysis due to the scatter of experimentally observed rate constants among different proteins.

A unified mechanism of protein folding: hierarchy

Figure 5 compares the non-two-state and the two-state proteins in terms of the dependence of the logarithmic rate constant on Q_d. Both the value of $\log(k_{UN}/Q_d)$ and its dependence on Q_d in the case of the two-state proteins were very similar to those found in the plots of $\log(k_I/Q_d)$ and $\log(k_N/Q_d)$ against Q_d for the non-two-state proteins. These similarities demonstrated clearly that

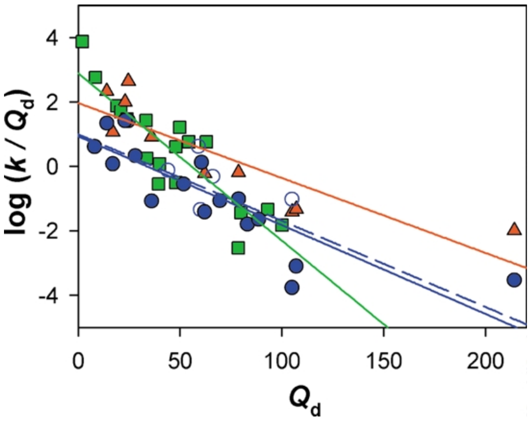

Figure 5. A comparison of the folding rates of non-two-state and two-state folders. The logarithmic rate constants are plotted against the number of sequence-distant native pairs. Filled squares (■) represent the folding of two-state folders. Filled triangles (▲) represent the formation of the intermediate of non-two-state folders. Filled circles (●) and open circles (○) represent the formation of the native state for the 17 protein and the 22 protein data set, respectively, of non-two-state folders. The green, red, and blue continuous lines represent the best linear fit for the formation of the native state for 18 two-state folders, the intermediate for ten non-two-state folders, and the native state for 17 non-two-state folders, respectively. The broken blue line represents the best linear fit for the formation of the native state for 22 non-two-state folders.

the mechanism of folding does not differ between the two classes of proteins. Similarities between the non-two-state and the two-state proteins were revealed in terms of the dependence on the other topological parameters, ACO and cliquishness. These results suggest that non-two-state folding, with the accumulation of a productive folding intermediate, may be a more common mechanism of protein folding; moreover, two-state folding is apparently a simplified version of the more common non-two-state folding.

The known structural characteristics of the folding intermediate in the case of the non-two-state proteins suggest strongly the hierarchy of the three-dimensional structure of a native protein as an important factor in determining the mechanism of protein folding. In general, the folding intermediate in many globular proteins has the following characteristics of the molten globule state:[1,3] (1) the presence of a substantial amount of native-like secondary structure; (2) the absence of most of the specific tertiary structure associated with the tight packing of side-chains; (3) the presence of a native-like backbone topology, and the compactness of the overall shape of the molecule, with a radius that is only 10–30% larger than that of the native state; and (4) heterogeneity of the three-dimensional structure, in which certain subdomains of the molecule are more organized than others.

The presence of the molten globule-like intermediate at an early stage of kinetic refolding, together with the productive nature of the intermediate evidenced by the present results, suggests a two-stage hierarchical model as a general mechanism of protein folding.[3,25] In this model, the protein folding process is divided into two stages: stage I, formation of the molten globule state (I) from the unfolded state (U); and stage II, formation of the native state (N) from the molten globule state, which can be depicted as follows:

$$\mathrm{U}\xrightarrow{\mathrm{I}}\mathrm{I}\xrightarrow{\mathrm{II}}\mathrm{N}$$

However, at least two questions arise with regard to the two-stage hierarchical model of folding. First, if the two-stage model is more general, why do certain small globular proteins exhibit two-state (i.e. single-stage) folding? Second, if stage II corresponds to the process of the specific tertiary packing of side-chains from the molten globule to the native state, why does the k_{N} for non-two-state folders show essentially the same dependence on native backbone topology (i.e. Q_{d}) as k_{I} (see Figure 5)?

As regards the first question as it relates to the two-state folders, there are two possible explanations. First, this effect could be due to the movement of the rate-limiting step from stage II in non-two-state folders to stage I in two-state folders.[26] When the size of a protein decreases, it becomes easier to determine the specific conformation of side-chain packing due to the decrease in the number of specific interactions, thereby making the first stage, rather than the second stage, rate-limiting. Thus, this process would lead to the two-state kinetic behavior of the protein. On the other hand, a second explanation would ascribe the two-state behavior to the destabilization of the folding intermediate. When the intermediate is less stable than the unfolded state, a protein must undergo simple two-state folding without any accumulation of the intermediate. If the first explanation is applied to the two-state type of folding, the rate constant k_{UN} of the two-state protein may coincide with the rate constant k_{I}, at which the intermediate forms in the non-two-state protein. On the other hand, if the second explanation is applied, with the rate-limiting step remaining at stage II, then the rate constant k_{UN} of the two-state protein may coincide with the rate constant k_{N}, at which the native state for the non-two-state protein is formed. Figure 5 shows that the k_{UN} values for the two-state proteins coincide with the k_{I} values for the non-two-state proteins at Q_{d} values of less than 25, suggesting the validity of the first explanation. However, when Q_{d} is larger than 25, the two-state k_{UN} shows variation in either the non-two-state k_{I} or k_{N}, or between the k_{I} and k_{N}, and hence both mechanisms given by the two explanations may play a role in the folding of two-state proteins. These conclusions are supported by a hidden intermediate model with partially folded intermediates behind the rate-limiting step proposed recently by Bai,[27] and a sequential folding pathway model with consecutive distinct barriers and a few obligatory high-energy intermediates proposed by Sanchez & Kiefhaber.[28]

However, it remains unclear why the k_N for non-two-state folders shows essentially the same dependence on Q_d as does the k_I (Figure 5). This question may be even more difficult to answer than those discussed above; however, the known structural characteristics of the folding intermediates indicate that stage II includes the process of the specific tertiary packing of side-chains. It is possible that this packing process is dominated by the organization of the native backbone topology. Thus, the folding that occurs in stage II would again be correlated with the native backbone topology, but would take place in a concerted manner together with the tertiary packing of side-chains and the organization of the backbone topology.

Methods

Numerical analysis

The calculations for the above structural parameters were performed using computer programs coded by our group (Compaq Visual Fortran Professional version 6.1.0 and Visual Basic version 6.0).

Correlation coefficient, *r*

The quality of the correlation was evaluated by Pearson's linear correlation coefficient, r, which is defined as:

$$r = \sum_i \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{4}$$

where the x_i and y_i are variables. We evaluated the quality of the correlation using Spearman's rank correlation coefficient,[29] and the results obtained from the rank correlation were essentially the same as those obtained from the linear correlation.

Partial correlation coefficient, *r**

Partial correlation coefficient, r^*, which represents the strength of the correlation between two variables (x and y) in which the effect of the third variable z is eliminated, is defined as:

$$r^* = \frac{r_{xy} - r_{yz} r_{zx}}{\sqrt{(1 - r_{yz}^2)(1 - r_{zx}^2)}} \tag{5}$$

where r_{ij} is the correlation coefficient between variables i and j.[29,30]

Testing the significance of the correlation

To test the significance of the correlation between the two variables, we estimated the 95% confidence interval of the Pearson's correlation coefficient using the Fisher Z transformation, in which the variance of the Z corresponding to the given r is $(N-3)^{-1}$.[30] For estimation of the partial correlation coefficient, the same procedure as described above was performed, but $(N-4)^{-1}$ was used as the variance of the Z distribution instead of $(N-3)^{-1}$.[29,30] The calculations were performed using computer programs coded by our group (Microsoft® EXCEL 2000 and Visual Basic version 6.0).

Estimation of the folding rate constants under the standard experimental condition

Because the experimental conditions (pH, temperature, and Na_2SO_4 concentration) are slightly different depending on the proteins listed in Tables 1 and 2, we have estimated the folding rate constants for the proteins under the standard condition (pH 7, 25 °C, and 0 M Na_2SO_4) by:

$$\log(k_{es}) = \log(k_{obs}) + \frac{1}{2}\Delta\mathrm{pH}\{-1,1\} + \frac{1}{15}\Delta T\{0,1\} - \frac{1}{0.4}[\mathrm{S}]\{0,1\} \tag{6}$$

where k_{es} and k_{obs} are the estimated rate constant under the standard condition and the experimentally observed rate constant, respectively, ΔpH is the absolute value of the difference in pH between the experimental and standard conditions, ΔT is the difference of T (°C) between the experimental and standard condition, [S] is the concentration of Na_2SO_4 (M), and $\{a,b\}$ is a random number between a and b. The prefactors were estimated from the experimental data of several proteins in which the pH-dependence, the T-dependence and the [S]-dependence of the folding rate constant were investigated, respectively.[31–34]

The prefactor, 1/2, of the second term means that a two-unit change in pH leads to a maximally tenfold difference in the folding rate constant, and similarly, the prefactor, 1/15, in the third term means that a 15 deg. C temperature decrease leads to a maximally tenfold decrease in the folding rate constant. Although these estimates were based on the rate constant data of folding into the native state, we assumed the same prefactor values for estimation of the $\log(k_I)$ values of the proteins. The calculations were performed using computer programs coded by our group (Visual Basic version 6.0).

Acknowledgements

This work was supported by Grants-in-Aid for Scientific Research on Priority Areas (no. 15076201) from the Ministry of Education, Culture, Science and Technology of Japan. The authors express their appreciation to Mr T. Ueki for his assistance in recalculating the structure-based parameters.

References

1. Ptitsyn, O. B. (1995). Molten globule and protein folding. *Advan. Protein Chem.* **47**, 83–229.
2. Kuwajima, K. (1989). The molten globule state as a clue for understanding the folding and cooperativity

of globular-protein structure. *Proteins: Struct. Funct. Genet.* **6**, 87–103.

3. Arai, M. & Kuwajima, K. (2000). Role of the molten globule state in protein folding. *Advan. Protein Chem.* **53**, 209–282.
4. Jackson, S. E. (1998). How do small single-domain proteins fold? *Fold. Des.* **3**, R81–R91.
5. Baldwin, R. L. (1995). The nature of protein folding pathways: the classical *versus* the new view. *J. Biomol. NMR*, **5**, 103–109.
6. Lazaridis, T. & Karplus, M. (1997). "New view" of protein folding reconciled with the old through multiple unfolding simulations. *Science*, **278**, 1928–1931.
7. Dill, K. A. & Chan, H. S. (1997). From Levinthal to pathways to funnels. *Nature Struct. Biol.* **4**, 10–19.
8. Dobson, C. M., Sali, A. & Karplus, M. (1998). Protein folding: a perspective from theory and experiment. *Angew. Chem. Int. Ed.* **37**, 868–893.
9. Plaxco, K. W., Simons, K. T. & Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985–994.
10. Plaxco, K. W., Simons, K. T., Ruczinski, I. & Baker, D. (2000). Topology, stability, sequence, and length: defining the determinants of two-state protein folding kinetics. *Biochemistry*, **39**, 11177–11183.
11. Galzitskaya, O. V., Garbuzynskiy, S. O., Ivankov, D. N. & Finkelstein, A. V. (2003). Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics. *Proteins: Struct. Funct. Genet.* **51**, 162–166.
12. Ivankov, D. N., Garbuzynskiy, S. O., Alm, E., Plaxco, K. W., Baker, D. & Finkelstein, A. V. (2003). Contact order revisited: influence of protein size on the folding rate. *Protein Sci.* **12**, 2057–2062.
13. Micheletti, C. (2003). Prediction of folding rates and transition-state placement from native-state geometry. *Proteins: Struct. Funct. Genet.* **51**, 74–84.
14. Kuwajima, K. & Arai, M. (2002). [Two views of protein folding: what is the universal view?]. *Tanpakushitsu Kakusan Koso*, **47**, 657–662.
15. Roder, H., Elove, G. A. & Shastry, M. C. R. (2000). Early stages of protein folding. In *Mechanisms of Protein Folding* (Pain, R. H., ed.), 2nd edit., pp. 65–104, Oxford University Press, New York.
16. Robinson, C. V. (2000). Protein folding monitored by mass spectroscopy. In *Mechanisms of Protein Folding* (Pain, R. H., ed.), 2nd edit., pp. 105–117, Oxford University Press, New York.
17. Kamagata, K., Sawano, Y., Tanokura, M. & Kuwajima, K. (2003). Multiple parallel-pathway folding of proline-free staphylococcal nuclease. *J. Mol. Biol.* **332**, 1143–1153.
18. Silow, M. & Oliveberg, M. (1997). High-energy channeling in protein folding. *Biochemistry*, **36**, 7633–7637.
19. Grantcharova, V., Alm, E. J., Baker, D. & Horwich, A. L. (2001). Mechanisms of protein folding. *Curr. Opin. Struct. Biol.* **11**, 70–82.
20. Makarov, D. E. & Plaxco, K. W. (2003). The topomer search model: a simple, quantitative theory of two-state protein folding kinetics. *Protein Sci.* **12**, 17–26.
21. Thirumalai, D. (1995). From minimal models to real proteins: time scales for protein folding kinetics. *J. Phys. I (France)*, **5**, 1457–1467.
22. Gutin, A. M., Abkevich, V. I. & Shakhnovich, E. I. (1996). Chain length scaling of protein folding time. *Phys. Rev. Letters*, **77**, 5433–5436.
23. Finkelstein, A. V. & Badretdinov, 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.
24. Koga, N. & 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.
25. Kuwajima, K. & Arai, M. (2000). The molten globule state: the physical picture and biological significance. In *Mechanisms of Protein Folding* (Pain, R. H., ed.), 2nd edit., pp. 138–174, Oxford University Press, New York.
26. Kuwajima, K. (2002). The role of the molten globule state in protein folding: the search for a universal view of folding. *Proc. Ind. Natl Sci. Acad.* **68**, 333–340.
27. Bai, Y. (2003). Hidden intermediates and levinthal paradox in the folding of small proteins. *Biochem. Biophys. Res. Commun.* **305**, 785–788.
28. Sanchez, I. E. & Kiefhaber, T. (2003). Evidence for sequential barriers and obligatory intermediates in apparent two-state protein folding. *J. Mol. Biol.* **325**, 367–376.
29. Thorndike, R. M. (1978). *Correlational Procedures for Research*, Gardner Press, New York.
30. Dunn, O. J. & Clark, V. A. (1987). *Applied Statistics: Analysis of Variance and Regression*, 2nd edit., Wiley, New York.
31. Finke, J. M. & Jennings, P. A. (2002). Interleukin-1β folding between pH 5 and 7: experimental evidence for three-state folding behavior and robust transition state positions late in folding. *Biochemistry*, **41**, 15056–15067.
32. Otzen, D. E. & Oliveberg, M. (1999). Salt-induced detour through compact regions of the protein folding landscape. *Proc. Natl Acad. Sci. USA*, **96**, 11746–11751.
33. Gorski, S. A., Capaldi, A. P., Kleanthous, C. & Radford, S. E. (2001). Acidic conditions stabilize intermediates populated during the folding of Im7 and Im9. *J. Mol. Biol.* **312**, 849–863.
34. Kuhlman, B., Luisi, D. L., Evans, P. A. & Raleigh, D. P. (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.
35. Capaldi, A. P., Shastry, M. C. R., Kleanthous, C., Roder, H. & Radford, S. E. (2001). Ultrarapid mixing experiments reveal that Im7 folds *via* an on-pathway intermediate. *Nature Struct. Biol.* **8**, 68–72.
36. Walkenhorst, W. F., Green, S. M. & Roder, H. (1997). Kinetic evidence for folding and unfolding intermediates in staphylococcal nuclease. *Biochemistry*, **36**, 5795–5805.
37. Teilum, K., Maki, K., Kragelund, B. B., Poulsen, F. M. & Roder, H. (2002). Early kinetic intermediate in the folding of acyl-CoA binding protein detected by fluorescence labeling and ultrarapid mixing. *Proc. Natl Acad. Sci. USA*, **99**, 9807–9812.
38. Chattopadhyay, K., Zhong, S., Yeh, S. R., Rousseau, D. L. & Frieden, C. (2002). The intestinal fatty acid binding protein: the role of turns in fast and slow folding processes. *Biochemistry*, **41**, 4040–4047.
39. Dalessio, P. M. & Ropson, I. J. (2000). β-sheet proteins with nearly identical structures have different folding intermediates. *Biochemistry*, **39**, 860–871.
40. Burns, L. L. & Ropson, I. J. (2001). Folding of intracellular retinol and retinoic acid binding proteins. *Proteins: Struct. Funct. Genet.* **43**, 292–302.

41. Park, S. H., Shastry, M. C. R. & Roder, H. (1999). Folding dynamics of the B1 domain of protein G explored by ultrarapid mixing. *Nature Struct. Biol.* **6**, 943–947.
42. Svensson, A. K., O'Neill, J. C., Jr & Matthews, C. R. (2003). The coordination of the isomerization of a conserved non-prolyl cis peptide bond with the rate-limiting steps in the folding of dihydrofolate reductase. *J. Mol. Biol.* **326**, 569–583.
43. Ikura, T. & Fersht, A. R. (2001). [Folding mechanism and folding rate]. *Tanpakushitsu Kakusan Koso*, **46**, 1553–1559.
44. Uzawa, T., Akiyama, S., Kimura, T., Takahashi, S., Ishimori, K., Morishima, I. & Fujisawa, T. (2004). Collapse and search dynamics of apomyoglobin folding revealed by submillisecond observations of α-helical content and compactmess. *Proc. Natl Acad. Sci. USA*, **101**, 1171–1176.
45. Parker, M. J., Dempsey, C. E., Lorch, M. & Clarke, A. R. (1997). Acquisition of native β-strand topology during the rapid collapse phase of protein folding. *Biochemistry*, **36**, 13396–13405.
46. Parker, M. J. & Marqusee, S. (1999). The cooperativity of burst phase reactions explored. *J. Mol. Biol.* **293**, 1195–1210.
47. Kern, G., Handel, T. & Marqusee, S. (1998). Characterization of a folding intermediate from HIV-1 ribonuclease H. *Protein Sci.* **7**, 2164–2174.
48. Raschke, T. M., Kho, J. & Marqusee, S. (1999). Confirmation of the hierarchical folding of RNase H: a protein engineering study. *Nature Struct. Biol.* **6**, 825–831.
49. Capaldi, A. P., Ferguson, S. J. & Radford, S. E. (1999). The Greek key protein apo-pseudoazurin folds through an obligate on-pathway intermediate. *J. Mol. Biol.* **286**, 1621–1632.
50. Laurents, D. V., Corrales, S., Elias-Arnanz, M., Sevilla, P., Rico, M. & Padmanabhan, S. (2000). Folding kinetics of phage 434 Cro protein. *Biochemistry*, **39**, 13963–13973.
51. Khan, F., Chuang, J. I., Gianni, S. & Fersht, A. R. (2003). The kinetic pathway of folding of barnase. *J. Mol. Biol.* **333**, 169–186.
52. Calloni, G., Taddei, N., Plaxco, K. W., Ramponi, G., Stefani, M. & Chiti, F. (2003). Comparison of the folding processes of distantly related proteins. Importance of hydrophobic content in folding. *J. Mol. Biol.* **330**, 577–591.
53. Parker, M. J., Spencer, J. & Clarke, A. R. (1995). An integrated kinetic analysis of intermediates and transition states in protein folding reactions. *J. Mol. Biol.* **253**, 771–786.
54. Cota, E. & Clarke, J. (2000). Folding of β-sandwich proteins: three-state transition of a fibronectin type III module. *Protein Sci.* **9**, 112–120.
55. Fowler, S. B. & Clarke, J. (2001). Mapping the folding pathway of an immunoglobulin domain: structural detail from φ value analysis and movement of the transition state. *Structure (Camb)*, **9**, 355–366.
56. Tang, K. S., Guralnick, B. J., Wang, W. K., Fersht, A. R. & Itzhaki, L. S. (1999). Stability and folding of the tumour suppressor protein p16. *J. Mol. Biol.* **285**, 1869–1886.
57. Munoz, V., Lopez, E. M., Jager, M. & Serrano, L. (1994). Kinetic characterization of the chemotactic protein from *Escherichia coli*, CheY. Kinetic analysis of the inverse hydrophobic effect. *Biochemistry*, **33**, 5858–5866.
58. Viguera, A. R., Martinez, J. C., Filimonov, V. V., Mateo, P. L. & Serrano, L. (1994). Thermodynamic and kinetic analysis of the SH3 domain of spectrin shows a two-state folding transition. *Biochemistry*, **33**, 2142–2150.
59. Grantcharova, V. P. & Baker, D. (1997). Folding dynamics of the src SH3 domain. *Biochemistry*, **36**, 15685–15692.
60. Guijarro, J. I., Morton, C. J., Plaxco, K. W., Campbell, I. D. & Dobson, C. M. (1998). Folding kinetics of the SH3 domain of PI3 kinase by real-time NMR combined with optical spectroscopy. *J. Mol. Biol.* **276**, 657–667.
61. Plaxco, K. W., Guijarro, J. I., Morton, C. J., Pitkeathly, M., Campbell, I. D. & Dobson, C. M. (1998). The folding kinetics and thermodynamics of the Fyn-SH3 domain. *Biochemistry*, **37**, 2529–2537.
62. Ferguson, N., Capaldi, A. P., James, R., Kleanthous, C. & Radford, S. E. (1999). Rapid folding with and without populated intermediates in the homologous four-helix proteins Im7 and Im9. *J. Mol. Biol.* **286**, 1597–1608.
63. Perl, D., Welker, C., Schindler, T., Schroder, K., Marahiel, M. A., Jaenicke, R. & Schmid, F. X. (1998). Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins. *Nature Struct. Biol.* **5**, 229–235.
64. Clarke, J., Hamill, S. J. & Johnson, C. M. (1997). Folding and stability of a fibronectin type III domain of human tenascin. *J. Mol. Biol.* **270**, 771–778.
65. Spector, S. & Raleigh, D. P. (1999). Submillisecond folding of the peripheral subunit-binding domain. *J. Mol. Biol.* **293**, 763–768.
66. Burton, R. E., Huang, G. S., Daugherty, M. A., Fullbright, P. W. & Oas, T. G. (1996). Microsecond protein folding through a compact transition state. *J. Mol. Biol.* **263**, 311–322.
67. Scalley, M. L., Yi, Q., Gu, H. D., McCormack, A., Yates, J. R., III & Baker, D. (1997). Kinetics of folding of the IgG binding domain of peptostreptoccocal protein L. *Biochemistry*, **36**, 3373–3382.
68. Villegas, V., Azuaga, A., Catasus, L., Reverter, D., Mateo, P. L., Aviles, F. X. & Serrano, L. (1995). Evidence for a two-state transition in the folding process of the activation domain of human procarboxypeptidase A2. *Biochemistry*, **34**, 15105–15110.
69. Jackson, S. E. & Fersht, A. R. (1991). Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition. *Biochemistry*, **30**, 10428–10435.
70. Van Nuland, N. A., Meijberg, W., Warner, J., Forge, V., Scheek, R. M., Robillard, G. T. & Dobson, C. M. (1998). Slow cooperative folding of a small globular protein HPr. *Biochemistry*, **37**, 622–637.
71. Reid, K. L., Rodriguez, H. M., Hillier, B. J. & Gregoret, L. M. (1998). Stability and folding properties of a model β-sheet protein, *Escherichia coli* CspA. *Protein Sci.* **7**, 470–479.
72. Guerois, R. & Serrano, L. (2000). The SH3-fold family: experimental evidence and prediction of variations in the folding pathways. *J. Mol. Biol.* **304**, 967–982.
73. Main, E. R. G., Fulton, K. F. & Jackson, S. E. (1999). Folding pathway of FKBP12 and characterisation of the transition state. *J. Mol. Biol.* **291**, 429–444.
74. van Nuland, N. A., Chiti, F., Taddei, N., Raugei, G., Ramponi, G. & Dobson, C. M. (1998). Slow folding of muscle acylphosphatase in the absence of intermediates. *J. Mol. Biol.* **283**, 883–891.
75. Kuhlman, B., Boice, J. A., Fairman, R. & Raleigh, D. P. (1998). Structure and stability of the N-terminal

domain of the ribosomal protein L9: evidence for rapid two-state folding. *Biochemistry*, **37**, 1025–1032.

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

Edited by A. R. Fersht

(Received 8 December 2003; received in revised form 1 April 2004; accepted 1 April 2004)

Note added in proof: A recent paper by Vu *et al.* (*Biochemistry* (2004), **43**, 3346–3356) suggests that the folding of barnase apparently follows a two-state reaction, although we have classified the protein as a non-two-state folder. Thus, we recalculated the correlations between the structure-based parameters and the rate constants of formation of the native state of the non-two-state folders by excluding barnase from the data set. The results indicated that the correlations obtained from the data without barnase were essentially the same as those obtained from the original data with barnase, indicating that any conclusions of the present paper were not affected by inclusion or exclusion of barnase. The r and ρ values for each correlation were as follows: $\log(k_N)$ with L ($r = -0.71$ (-0.66); $-0.89 \le \rho \le -0.32$ ($-0.85 \le \rho \le -0.32$)), $\log(k_N)$ with L^{ν} ($\nu = 0.1$ to 1) ($r = -0.72$ to -0.71 (-0.67 to -0.66)), $\log(k_N)$ with $\log(L)$ ($r = -0.72$ (-0.66); $-0.89 \le \rho \le -0.34$ ($-0.85 \le \rho \le -0.33$), and its slope, -6.1 ± 1.6 (-5.5 ± 1.4)), $\log(k_N)$ with *ACO* ($r = -0.71$ (-0.67); $-0.89 \le \rho \le -0.34$ ($-0.85 \le \rho \le -0.33$)), $\log(k_N)$ with *RCO* ($r = 0.12$ (0.08); $-0.40 \le \rho \le 0.58$ ($-0.36 \le \rho \le 0.50$)), $\log(k_N)$ with $Q_{l=2}$ ($r = -0.69$ (-0.67); $-0.89 \le \rho \le -0.30$ ($-0.85 \le \rho \le -0.33$)), $\log(k_N)$ with Q_d ($r = -0.79$ (-0.74); $-0.92 \le \rho \le -0.48$ ($-0.89 \le \rho \le -0.45$)), $\log(k_N/Q_d)$ with Q_d ($r = -0.84$ (-0.82); $-0.94 \le \rho \le -0.60$ ($-0.92 \le \rho \le -0.59$)), and the partial correlation of Q_d with $\log(k_N/Q_d)$ after eliminating the effect of L ($r^* = -0.66$ (-0.64); $-0.87 \le \rho^* \le -0.22$ ($-0.84 \le \rho^* \le -0.27$)), whereby the values outside and inside of the parentheses represent the values for the 16 protein and 21 protein data sets, respectively.