Effect of Finite Size on Cooperativity and Rates of Protein Folding[†]

Maksim Kouza,[‡] Mai Suan Li,[‡] Edward P. O'Brien, Jr.,[§] Chin-Kun Hu,[∥,⊥] and D. Thirumalai*,[§,#]

Institute of Physics, Polish Academy of Sciences, Al. Lotnikow 32/46, 02-668 Warsaw, Poland, Biophysics Program, Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742, Institute of Physics, Academia Sinica, Nankang, Taipei 11529, Taiwan, National Center for Theoretical Sciences at Taipei, Physics Division, National Taiwan University, Taipei 10617, Taiwan, and Department of Chemistry and Biochemistry, University of Maryland, College Park, Maryland 20742

Received: July 8, 2005; In Final Form: September 30, 2005

We analyze the dependence of cooperativity of the thermal denaturation transition and folding rates of globular proteins on the number of amino acid residues, N, using lattice models with side chains, off-lattice Go models, and the available experimental data. A dimensionless measure of cooperativity, Ω_c ($0 < \Omega_c < \infty$), scales as $\Omega_c \approx N^{\zeta}$. The results of simulations and the analysis of experimental data further confirm the earlier prediction that ζ is universal with $\zeta = 1 + \gamma$, where exponent γ characterizes the susceptibility of a self-avoiding walk. This finding suggests that the structural characteristics in the denaturated state are manifested in the folding cooperativity at the transition temperature. The folding rates k_F for the Go models and a dataset of 69 proteins can be fit using $k_F = k_F^0 \exp(-cN^{\beta})$. Both $\beta = {}^1/_2$ and ${}^2/_3$ provide a good fit of the data. We find that $k_F = k_F^0 \exp(-cN^{1/2})$, with the average (over the dataset of proteins) $k_F^0 \approx (0.2\ \mu s)^{-1}$ and $c \approx 1.1$, can be used to estimate folding rates to within an order of magnitude in most cases. The minimal models give identical N dependence with $c \approx 1$. The prefactor for off-lattice Go models is nearly 4 orders of magnitude larger than the experimental value.

I. Introduction

Single-domain globular proteins are mesoscopic systems that self-assemble, under folding conditions, to a compact state with definite topology. Given that the folded states of proteins are only on the order of tens of angstroms (the radius of gyration $R_g \approx 3N^{1/3}$ Å,[1] where N is the number of amino acids), it is surprising that they undergo highly cooperative transitions from an ensemble of unfolded states to the native state.[2,3] Similarly, there is a wide spread in the folding times as well.[4−6] The rates of folding vary by nearly nine orders of magnitude. Sometime ago, it was shown theoretically that the folding time, τ_F, should depend on N,[7−9] but only recently has experimental data confirmed this prediction.[4,6,10−12] It has been shown that τ_F can be approximately evaluated using $\tau_F \approx \tau_F^0 \exp(N^{\beta})$ where ${}^1/_2 \leq \beta < {}^2/_3$ with the prefactor τ_F^0 being on the order of a microsecond.

Much less attention has been paid to finite size effects on the cooperativity of transition from unfolded states to the native basin of attraction (NBA). Because N is finite, large conformational fluctuations are possible, which require careful examination.[10,13−15] For large enough N, it is likely that the folding or melting temperature itself may not be unique.[16−18] Although substantial variations in T_m are unlikely, it has already been shown that there is a range of temperatures over which individual residues in a protein achieve their native state ordering.[16] On the other hand, the global cooperativity, as measured by the dimensionless parameter Ω_c (see below for definition) has been shown to scale as[14]

$$\Omega_c \approx N^{\zeta} \tag{1}$$

The surprising finding in eq 1 requires some discussion. The result in eq 1 is obtained using the following arguments. From the definition of Ω_c (see eq 2 and subsequent discussions), it follows that $f_N \propto P_{NBA}$ where P_{NBA} is the probability of being in the native basin of attraction. Experimentally, P_{NBA} (or an equivalent measure) is assessed using spectroscopic signatures (circular dichroism (CD), fluorescence, etc.) of proteins at low temperatures or at low denaturant concentrations. In computations, P_{NBA} is computed from the temperature dependence of the structural overlap function (eq 6). The fraction of molecules in the native state is $f_N = 1 - \langle\chi\rangle$ where $\langle\ \rangle$ is the thermal average. Thus, the dimensionless measure of cooperativity Ω_c can be written as

$$\Omega_c = \frac{T_F^2}{\Delta T}\frac{\partial\langle\chi\rangle}{\partial T} \tag{2}$$

where T_F is the folding temperature and ΔT is the full width at half-maximum of $d\langle\chi\rangle/dT$. The folding temperature can be identified with the peak in $d\langle\chi\rangle/dT$ or in the fluctuations in χ, namely, $\Delta\chi = \langle\chi^2\rangle - \langle\chi\rangle^2$. Using an analogy to magnetic systems, we identify $T(\partial\langle\chi\rangle/\partial h) = \Delta\chi$ where h is an "ordering field" that is conjugate to χ. Since $\Delta\chi$ is dimensionless, we expect $h \approx T$ for proteins, and hence, $T(\partial\langle\chi\rangle/\partial T)$ is like susceptibility. Hence, the scaling of Ω_c on N should follow the way $(T_F/\Delta T)\ \Delta\chi$ changes with N.

With the analogy to magnetic systems, we can obtain eq 1 by noting that for efficient folding in proteins $T_F \approx T_{\Theta}$ where

[†] Part of the special issue "Donald G. Truhlar Festschrift".
[‡] Polish Academy of Sciences.
[§] Institute for Physical Science and Technology, University of Maryland.
[∥] Academia Sinica.
[⊥] National Taiwan University.
[#] Department of Chemistry and Biochemistry, University of Maryland.

10.1021/jp053770b CCC: $33.50
Published on Web 12/09/2005

T_Θ is the temperature at which the coil-to-globule transition occurs. It has been argued that T_F for proteins may well be a tricritical point, because the transition at T_F is first-order while the collapse transition is (typically) second-order. Then, as temperature approaches from above, we expect that the characteristics of polypeptide chains at T_Θ should manifest themselves in the folding cooperativity. At or above T_F, the global conformations of the polypeptide chains as measured by R_g obey the Flory law, i.e., $R_g \approx aN^\nu$ where $\nu \approx 0.6$.[19]

We expect that $R_g \approx \Delta T^{-\nu}$ at temperatures close to $T_F \approx T_\theta$. If we use the magnet analogy for random coils or self-avoiding walks, R_g is like the correlation length. Thus, $\Delta T/T_F \approx 1/N$. Because $T(\partial\langle\chi\rangle/\partial T)$ is a generalized susceptibility, we expect $T(\partial\langle\chi\rangle/\partial T) \approx N^\nu$. By combining these two reasons, eq 1 is obtained.

The expectation that the random coil nature of the polypeptide chains at $T \approx T_\theta \approx T_F$ should be reflected in the thermodynamic folding cooperativity has important consequences. First, the exponent ζ is universal and is typically related to γ. The exact mapping of the self-avoiding walk to n-vector spins or n-component field theory lets us use the numerical value of $\gamma \approx 1.2$ to get $\zeta = 2.2$. It should be recalled that entropy of the random coil states (or compact states) is extensive. However, because globally unfolded chains are like random coils ($R_g \approx aN^{0.6}$), there is a logarithmic correction to the entropy that is characterized by the γ exponent. It is for this reason that we predict $\Omega_c \approx N^\zeta$ with $\zeta = 2.2$ for temperatures in the vicinity of T_F.

In this paper, we use lattice models with side chains (LMSC), off-lattice Go models for 23 proteins, and experimental results for a number of proteins to further confirm the theoretical predictions. Our results show that $\zeta \approx 2.22$, which is *distinct from the expected result* ($\zeta = 2.0$) *for a strong first-order transition*.[20] The larger data set of proteins for which folding rates are available shows that the folding time scales as

$$\tau_F = \tau_0 \exp(cN^\beta) \tag{3}$$

with $c \approx 1.1$, $\beta = {}^1/_2$, and $\tau_0 \approx 0.2$ μs.

II. Models and Methods

A. Lattice Models with Side Chains (LMSC). Each amino acid is represented using the backbone (B) C_α atom that is covalently linked to a unified atom representing the side chain (SC). Both the C_α atoms and the SCs are confined to the vertexes of a cubic lattice with spacing a. Thus, a polypeptide chain consisting of N residues is represented using $2N$ beads. The energy of a conformation is

$$E = \epsilon_{bb}\sum_{i=1,j>i+1}^{N}\delta_{r_{ij}^{bb},a} + \epsilon_{bs}\sum_{i=1,j\neq i}^{N}\delta_{r_{ij}^{bs},a} + \epsilon_{ss}\sum_{i=1,j>i}^{N}\delta_{r_{ij}^{ss},a} \tag{4}$$

where ϵ_{bb}, ϵ_{bs}, and ϵ_{ss} are backbone−backbone (BB−BB), backbone−side chain (BB−SC), and side chain−side chain (SC−SC) contact energies, respectively. The distances r_{ij}^{bb}, r_{ij}^{bs}, and r_{ij}^{ss} are between BB, BS, and SS beads, respectively. The contact energies ϵ_{bb}, ϵ_{bs}, and ϵ_{ss} are taken to be −1 (in units of k_bT) for native and 0 for non-native interactions. The neglect of interactions between residues not present in the native state is the approximation used in the Go model. Because we are interested in general scaling behavior, the use of the Go model is justified. We should emphasize that Ω_c can also be used even for proteins that are not judged to be calorimetrically two-state like. Indeed, in the original study,[27] Ω_c was used to analyze pH-dependent cooperativity of the folding transition of apomyoglobin.

Our purpose here is to restrict ourselves to apparent two-state folders, and hence, we have used Go models. We expect our results to hold even for well-optimized sequences that also include non-native interactions.

B. Off-Lattice Model. We employ coarse-grained off-lattice models for polypeptide chains in which each amino acid is represented using only the C_α atoms.[21] Furthermore, we use a Go model[22] in which the interactions between residues forming native contacts are assumed to be attractive and the non-native interactions are repulsive. Thus, by definition for the Go model, the PDB structure is the native structure with the lowest energy. The energy of a conformation of the polypeptide chain specified by the coordinates r_i of the C_α atoms is[23]

$$E = \sum_{\text{bonds}} K_r(r_{i,i+1} - r_{0i,i+1})^2 + \sum_{\text{angles}} K_\theta(\theta_i - \theta_{0i})^2 + \sum_{\text{dihedral}} \{K_\phi^{(1)}[1 - \cos(\Delta\phi_i)] + K_\phi^{(3)}[1 - \cos 3(\Delta\phi_i)]\} + \sum_{i>j-3}^{\text{NC}} \epsilon_H[5R_{ij}^{12} - 6R_{ij}^{10}] + \sum_{i>j-3}^{\text{NNC}} \epsilon_H\left(\frac{C}{r_{ij}}\right)^{12} \tag{5}$$

Here, $\Delta\phi_i = \phi_i - \phi_{0i}$, $R_{ij} = r_{0ij}/r_{ij}$; $r_{i,i+1}$ is the distance between beads i and $i + 1$, θ_i is the bond angle between bonds ($i - 1$) and i, and ϕ_i is the dihedral angle around the ith bond and r_{ij} is the distance between the ith and jth residues. Subscripts 0, NC, and NNC refer to the native conformation, native contacts, and non-native contacts, respectively. Residues i and j are in native contact if r_{0ij} is less than a cutoff distance $d_c = 6$ Å, where r_{0ij} is the distance between the residues in the native conformation.

The first harmonic term in eq 5 accounts for chain connectivity, and the second term represents the bond angle potential. The potential for the dihedral angle degrees of freedom is given by the third term in eq 5. The interaction energy between residues that are separated by at least three beads is given by 10−12 Lennard-Jones potential. A soft sphere (last term in eq 5) repulsive potential disfavors the formation of non-native contacts. We choose $K_r = 100\epsilon_H/\text{Å}^2$, $K_\theta = 20\epsilon_H/\text{rad}^2$, $K_\phi^{(1)} = \epsilon_H$, and $K_\phi^{(3)} = 0.5\epsilon_H$, where ϵ_H is the characteristic hydrogen bond energy and $C = 4$ Å.

C. Simulations. For the LMSC, we performed Monte Carlo simulations using the previously well-tested move set MS3.[36] This move set ensures that ergodicity is obtained efficiently even for $N = 50$; it uses single, double, and triple bead moves.[38] Following standard practice, the thermodynamic properties are computed using the multiple histogram method.[25] The kinetic simulations are carried out by a quench from high temperature to a temperature at which the NBA is preferentially populated. The folding times are calculated from the distribution of first passage times.

For off-lattice models, we assume that the dynamics of the polypeptide chain obeys the Langevin equation. The equations of motion were integrated using the velocity form of the Verlet algorithm with the time step $\Delta t = 0.005\tau_L$, where $\tau_L = (ma^2/\epsilon_H)^{1/2} \approx 3$ ps. To calculate the thermodynamic quantities, we collected histograms for the energy and native contacts at five or six different temperatures (at each temperature, 20−50 trajectories were generated depending on proteins). As with the LMSC, we used the multiple histogram method[25] to obtain the thermodynamic parameters at all temperatures.

For off-lattice models, the probability of being in the native state is computed using

$$f = \frac{1}{Q_T} \sum_{i<j+1}^{N} \theta(1.2r_{0ij} - r_{ij})\Delta_{ij} \tag{6}$$

where Δ_{ij} is equal to 1 if residues i and j form a native contact and 0 otherwise, Q_T is the total number of native contacts, and $\theta(x)$ is the Heaviside function. For the LMSC model, we used the structural overlap function[24]

$$\chi = \frac{1}{2N^2 - 3N + 1}\Big[\sum_{i<j} \delta(r_{ij}^{ss} - r_{ij}^{ss,N}) + \sum_{i<j+1} \delta(r_{ij}^{bb} - r_{ij}^{bb,N}) + \sum_{i\neq j} \delta(r_{ij}^{bs} - r_{ij}^{bs,N})\Big] \tag{7}$$

The overlap function χ, which is one if the conformation of the polypeptide chain coincides with the native structure and is small for unfolded conformations, is an order parameter for the folding−unfolding transition. The probability of being in the native state f_N is $f_N = \langle f \rangle = 1 - \langle \chi \rangle$, where $\langle ... \rangle$ denotes a thermal average.

D. Cooperativity. The extent of cooperativity of the transition to the NBA from the ensemble of unfolded states is measured using the dimensionless parameter

$$\Omega_c = \frac{T_F^2}{\Delta T}\left|\frac{df_N}{dT}\right|_{T=T_F} \tag{8}$$

where ΔT is the full width at half-maximum of df_N/dT and the folding temperature T_F is identified with the maximum of df_N/dT. Two points about Ω_c are noteworthy. (1) For proteins that melt by a two-state transition, it is trivial to show that $\Delta H_{vH} = 4k_B \Delta T \Omega_c$, where ΔH_{vH} is the van't Hoff enthalpy at T_F. For an infinitely sharp two-state transition, there is a latent heat release at T_F, at which C_p can be approximated by a Δ function. In this case, $\Omega_c \rightarrow \infty$, which implies that ΔH_{vH} and the calorimetric enthalpy ΔH_{cal} (obtained by integrating the temperature dependence of the specific heat C_p) would coincide. It is logical to infer that as Ω_c increases the ratio $\kappa = \Delta H_{vH}/\Delta H_{cal}$ should approach unity. (2) Even for moderately sized proteins that undergo a two-state transition, $\kappa \approx 1$.[3] It is known that the extent of cooperativity depends on external conditions, as has been demonstrated for thermal denaturation of CI2 at several values of pH.[26] The values of κ for all pH values are ~1. However, the variation in cooperativity of CI2 as pH varies are reflected in the changes in Ω_c.[27] Therefore, we believe that Ω_c, which varies in the range $0 < \Omega_c < \infty$, is a better descriptor of the extent of cooperativity than κ. The latter merely tests the applicability of the two-state approximation.

III. Results

A. Dependence of Ω_c on N. For the 23 Go proteins listed in Table 1, we calculated Ω_c from the temperature dependence of f_N. In Figure 1, we compare the temperature dependence of $f_N(T)$ and $df_N(T)/dT$ for β-hairpin ($N = 16$) and *Bacillus subtilis* (CpsB, $N = 67$). It is clear that the transition width and the amplitudes of df_N/dT obtained using Go models compare only qualitatively well with experiments. As pointed out by Kaya and Chan,[28−31] the simple Go-like models consistently underestimate the extent of cooperativity. Nevertheless, both the models and the experiments show that Ω_c increases dramatically as N increases (Figure 1).

TABLE 1: List of 23 Proteins Used in the Simulations

protein	N	PDB code[a]	Ω_c[b]	$\delta\Omega_c$[c]
β-hairpin	16	1PGB	2.29	0.02
α-helix	21	no code	0.803	0.002
WW domain	34	1PIN	3.79	0.02
Villin headpiece	36	1VII	3.51	0.01
YAP65	40	1K5R	3.63	0.05
E3BD	45		7.21	0.05
hbSBD	52	1ZWV	51.4	0.2
protein G	56	1PGB	16.98	0.89
SH3 domain (α-spectrum)	57	1SHG	74.03	1.35
SH3 domain (fyn)	59	1SHF	103.95	5.06
IgG-binding domain of streptococcal protein L	63	1HZ6	21.18	0.39
chymotrypsin inhibitor 2 (CI-2)	65	2CI2	33.23	1.66
CspB (*Bacillus subtilis*)	67	1CSP	66.87	2.18
CspA	69	1MJC	117.23	13.33
ubiquitin	76	1UBQ	117.8	11.1
activation domain procarboxypeptidase A2	80	1AYE	73.7	3.1
His-containing phosphocarrier protein	85	1POH	74.52	4.2
hbLBD	87	1K8M	15.8	0.2
tenascin (short form)	89	1TEN	39.11	1.14
Twitchin Ig repeat 27	89	1TIT	44.85	0.66
S6	97	1RIS	48.69	1.31
FKBP12	107	1FKB	95.52	3.85
ribonuclease A	124	1A5P	69.05	2.84

[a] The native state for use in the Go model is obtained from the structures deposited in the Protein Data Bank. [b] Ω_c is calculated using eq 8 with $f_N = \langle \chi(T) \rangle$. [c] $2\,\delta\Omega_c = |\Omega_c - \Omega_{c1}| + |\Omega_c - \Omega_{c2}|$, where Ω_{c1} and Ω_{c2} are values of the cooperativity measure obtained by retaining only one-half the conformations used to compute Ω_c.

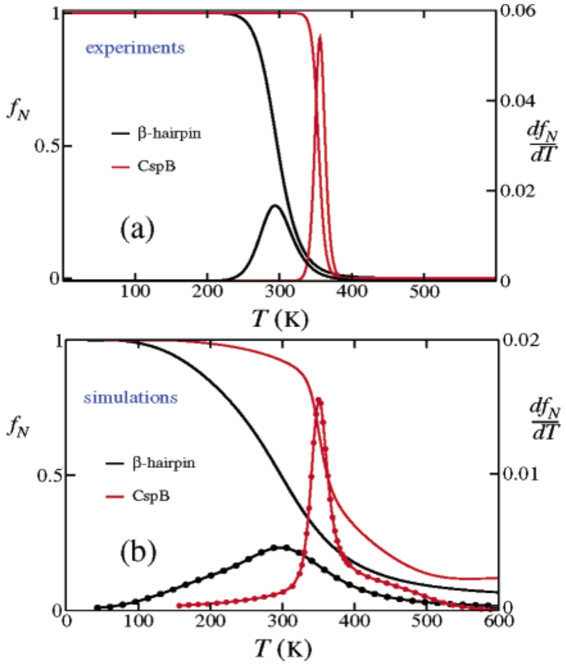

Figure 1. The temperature dependence of f_N and df_N/dT for β-hairpin ($N = 16$) and CpsB ($N = 67$). The scale for df_N/dT is given on the right. (a) The experimental curves were obtained by using $\Delta H = 11.6$ kcal/mol and $T_m = 297$ K, and $\Delta H = 54.4$ kcal/mol and $T_m = 354.5$ K for β-hairpin and CpsB, respectively. (b) The simulation results were calculated from $f_N = \langle \chi(T) \rangle$. The Go model gives only a qualitatively reliable measure of $f_N(T)$.

The variation of Ω_c with N for the 23 proteins obtained from the simulations of Go models is given in Figure 2. From the ln Ω_c−ln N plot, we obtain $\zeta = 2.40 \pm 0.20$ and $\zeta = 2.35 \pm 0.07$ for off-lattice models and LMSC, respectively. These values of ζ deviate from the theoretical prediction $\zeta \approx 2.22$. We suspect

TABLE 2: List of 34 Proteins for Which Ω_c Is Calculated Using Experimental Data[a]

protein	N	Ω_c[b]	$\delta\Omega_c$[c]	protein	N	Ω_c[b]	$\delta\Omega_c$[c]
BH8 β-hairpin[41]	12	12.9	0.5	SS07d[51]	64	555.2	56.2
HP1 β-hairpin[42]	15	8.9	0.1	CI2[26]	65	691.2	17.0
MrH3a β-hairpin[41]	16	54.1	6.2	CspTm[52]	66	558.2	56.3
β-hairpin[43]	16	33.8	7.4	Btk SH3[53]	67	316.4	25.9
Trp-cage protein[44]	20	24.8	5.1	binary pattern protein[54]	74	273.9	30.5
α-helix[45]	21	23.5	7.9	ADA2h[55]	80	332.0	35.2
villin headpeace[46]	35	112.2	9.6	hbLBD[56]	87	903.1	11.1
FBP28 WW domain[47 d]	37	107.1	8.9	tenascin Fn3 domain[57]	91	842.4	56.6
FBP28 W30A WW domain[47 d]	37	90.4	8.8	Sa RNase[58]	96	1651.1	166.6
WW prototype[47 d]	38	93.8	8.4	Sa3 RNase[58]	97	852.7	86.0
YAP WW[47 d]	40	96.9	18.5	HPr[59]	98	975.6	61.9
BBL[48]	47	128.2	18.0	Sa2 RNase[58]	99	1535.0	156.9
PSBD domain[48]	47	282.8	24.0	barnase[60]	110	2860.1	286.0
PSBD domain[48]	50	176.2	13.0	RNase A[61]	125	3038.5	42.6
hbSBD[49]	52	71.8	6.3	RNase B[61]	125	3038.4	87.5
B1 domain of protein G[50]	56	525.7	12.5	lysozyme[62]	129	1014.1	187.3
B2 domain of protein G[50]	56	468.4	20.0	interleukin-1β[63]	153	1189.6	128.6

[a] The calculated Ω_c values from experiments are significantly larger than those obtained using the Go models (see Table 1). [b] Ω_c is computed at $T = T_F = T_m$ using the experimental values of ΔH and T_m. [c] The error in $\delta\Omega_c$ is computed using the proceedure given in refs 14 and 35. [d] Data are averaged over two salt conditions at pH 7.0.

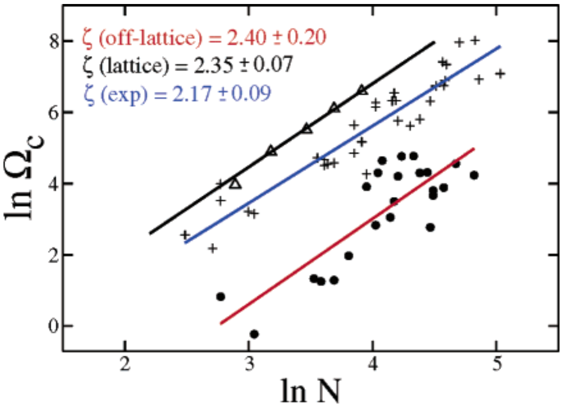

Figure 2. Plot of ln Ω_c as a function of ln N. The red line is a fit to the simulation data for the 23 off-lattice Go proteins from which we estimate $\zeta = 2.40 \pm 0.20$. The black line is a fit to the lattice models with side chains (N = 18, 24, 32, 40, and 50) with $\zeta = 2.35 \pm 0.07$. The blue line is a fit to the experimental values of Ω_c for 34 proteins (Table 2) with $\zeta = 2.17 \pm 0.09$. The larger deviation in ζ for the minimal models is due to lack of all the interactions that stabilize the native state.

that this is due to large fluctuations in the native state of polypeptide chains that are represented using minimal models. Nevertheless, the results for the minimal models rule out the value of $\zeta = 2$ that is predicted for systems that undergo first-order transition. The near-coincidence of ζ for both models show that the details of interactions are not relevant.

For the 34 proteins (Table 2) for which we could find thermal denaturation data, we calculated Ω_c using ΔH and T_F (referred to as the melting temperature T_m in the experimental literature). From the plot of ln Ω_c versus ln N, we find that $\zeta = 2.17 \pm 0.09$. The experimental value of ζ, which also deviates from $\zeta = 2$, is in much better agreement with the theoretical prediction. The analysis of the experimental data requires care, because the compiled results were obtained from a number of different laboratories around the world. Each laboratory uses different methods to analyze the raw experimental data, which invariably leads to varying methods to estimate errors in ΔH and T_m. To estimate the error bar for ζ, it is important to consider the errors in the computation of Ω_c. Using the reported experimental errors in T_m and ΔH, we calculated the variance $\delta^2\Omega_c$ using the standard expression for the error propagation.[14,39] The upper bound in the error in Ω_c for the 34 proteins is given in Table 2. To provide an accurate evaluation of the errors in the exponent ζ, we used a weighted linear fit, in which each value of ln Ω_c contributes to the fit with the weight proportional to its standard deviation.[14,39]

B. Dependence of Folding Free Energy Barrier on *N*. The simultaneous presence of stabilizing (between hydrophobic residues) and destabilizing interactions involving polar and charged residues in polypeptide chain renders the native state only marginally stable.[2] The hydrophobic residues enable the formation of compact structures, while polar and charged residues, for whom water is a good solvent, are better accommodated by extended conformations. Thus, in the folded state, the average energy gain per residue (compared to expanded states) is $-\epsilon_H$ ($\approx 1-2$ kcal/mol), whereas because of chain connectivity and surface area burial, the loss in free energy of exposed residues is $\epsilon_P \approx \epsilon_H$. Because there are a large number of solvent-mediated interactions that stabilize the native state, even when N is small, it follows from the central limit theorem that the barrier height $\beta\Delta G^{\ddagger}$, whose lower bound is the stabilizing free energy, should scale as $\Delta G^{\ddagger} \approx k_BT\sqrt{N}$.[7] A different physical picture has been used to argue that $\Delta G^{\ddagger} \approx k_BTN^{2/3}$.[8,9] Both scenarios show that the barrier to folding rates scales sublinearly with N.

The dependence of ln k_F ($k_F = \tau_F^{-1}$) on N using experimental data for 69 proteins[12] and the simulation results for the 23 proteins is consistent with the predicted behavior that $\Delta G^{\ddagger} = ck_BT\sqrt{N}$ with $c \approx 1$ (Figure 3). The correlation between the experimental results and the theoretical fit is 0.74, which is similar to the previous analysis using a set of 57 proteins.[10] It should be noted that the data can also be fit using $\Delta G^{\ddagger} \approx k_BTN^{2/3}$. The prefactor τ_F^0 using the $N^{2/3}$ fit is over an order of magnitude larger than for the $N^{1/2}$ behavior. In the absence of accurate measurements for a larger data set of proteins, it is difficult to distinguish between the two power laws for $\Delta G^{\ddagger}$.

Previous studies[32,33] have shown that there is a correlation between folding rates and Z-score, which can be defined as

$$Z_G = \frac{G_N - \langle G_U \rangle}{\sigma} \tag{9}$$

where G_N is the free energy of the native state, $\langle G_U \rangle$ is the average free energy of the unfolded states, and σ is the dispersion in the free energy of the unfolded states. From the

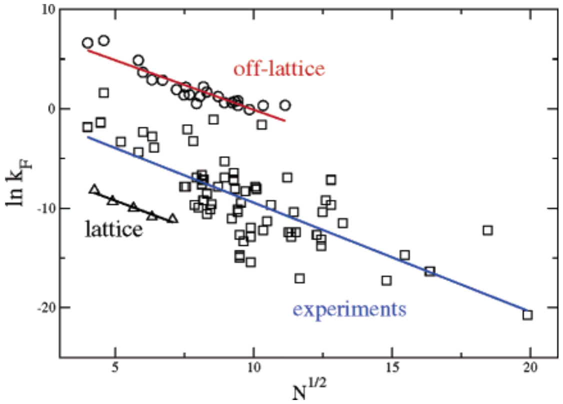

Figure 3. Folding rate of 69 real proteins (squares) is plotted as a function of $N^{1/2}$ (the straight line represents the fit $y = 1.54 - 1.10x$ with the correlation coefficient $R = 0.74$). The open circles represent the data obtained for 23 off-lattice Go proteins (see Table 1) (the linear fit $y = 9.84 - x$ and $R = 0.92$). The triangles denote the data obtained for lattice models with side chains ($N = 18$, 24, 32, 40, and 50; the linear fit $y = -4.01 - 1.1x$ and $R = 0.98$). For real proteins and off-lattice Go proteins, k_F is measured in μs^{-1}, whereas for the lattice models, it is measured in MCS^{-1} where MCS is Monte Carlo steps.

fluctuation formula, it follows that $\sigma = \sqrt{k_B T^2 C_p}$ so that

$$Z_G = \frac{\Delta G}{\sqrt{k_B T^2 C_p}} \qquad (10)$$

Since ΔG and C_p are extensive, it follows that $Z_G \approx N^{1/2}$. This observation establishes an intrinsic connection between the thermodynamics and kinetics of protein folding that involves formation and rearrangement of noncovalent interactions. In an interesting recent note,[12] it has been argued that the finding $\Delta G^{\ddagger} \approx k_B T\sqrt{N}$ can be interpreted in terms of n_σ in which ΔG in eq 10 is replaced by ΔH. In either case, there appears to be a thermodynamic rationale for the sublinear scaling of the folding free energy barrier.

IV. Conclusions

We have reexamined the dependence of the extent of cooperativity as a function of N using lattice models with side chains, off-lattice models, and experimental data on thermal denaturation. The finding that $\Omega_c \approx N^{\zeta}$ at $T \approx T_F$ with $\zeta > 2$ provides additional support for the earlier theoretical predictions.[14] More importantly, the present work also shows that the theoretical value for ζ is independent of the precise model used which implies that ζ is universal. It is surprising to find such general characteristics for proteins for which specificity is often an important property. We should note that accurate values of ζ and Ω_c can only be obtained using more refined models that perhaps include desolvation penalty.[29,34]

In accord with a number of theoretical predictions,[7−9,35−37] we found that the folding free energy barrier scales only sublinearly with N. The relatively small barrier is in accord with the marginal stability of proteins. Since the barriers to global unfolding are relatively small, it follows that there must be large conformational fluctuations even when the protein is in the NBA. Indeed, recent experiments show that such dynamical fluctuations that are localized in various regions of a monomeric protein might play an important functional role. These observations suggest that small barriers in proteins and RNA[40] might be general characteristics of all natural sequences.

There have been successful empirical attempts to obtain folding rates using simple characteristics of the native structures.[64−66] These studies show that proteins dominated by long-range (measured in terms of sequence separation) contacts between residues fold more slowly than those that have a large number of short-range contacts. The topological characteristics that are used in the contact order are not unrelated to size. Indeed, more recent considerations that take N into account give a better correlation between folding rates and modified contact order. Thus, both the architecture of the fold and size are important determinants of folding rate.

Acknowledgment. This work was supported in part by a KBN grant no. 1P03B01827, the National Science Foundation grant (CHE 05-14056), National Science Council in Taiwan under grant no. NSC 93-2112-M001-027, Academia Sinica (Taiwan) Grant AS-92-TP-A09, and National Science Council in Taiwan under grant no. NSC 94-2119-M002-001 (to C.K.H.).

References and Notes

(1) Dima, R. I.; Thirumalai, D. *J. Phys. Chem. B* **2004**, *108*, 6564−6570.

(2) (a) Poland, D.; Scheraga, H. A. *Theory of helix-coil transitions in biopolymers*; Academic Press: New York, 1970. (b) Creighton, T. E. *Proteins: Structures and Molecular Principles*; W. H. Freeman & Co.: New York, 1993.

(3) Privalov, P. L. *Adv. Phys. Chem.* **1979**, *33*, 167.

(4) Galzitskaya, O. V.; Garbuzynskiy, S. O.; Ivankov, D. N.; Finkelstein, A. V. *Proteins: Struct., Funct., Genet.* **2003**, *51*, 162−166.

(5) Ivankov, D. N.; Garbuzynskiy, S. O.; Alm, E.; Plaxco, K. W.; Baker, D.; Finkelstein, A. V. *Protein Sci.* **2003**, *12*, 2057−2062.

(6) Ivankov, D. N.; Finkelstein, A. V. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 8942−8944.

(7) Thirumalai, D. *J. Phys. (Paris)* **1995**, *5*, 1457−1467.

(8) Finkelstein, A. V.; Badretdinov, A. Ya. *Folding Des.* **1997**, *2*, 115−121.

(9) Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 6170−6175.

(10) Li, M. S.; Klimov, D. K.; Thirumalai, D. *Polymer* **2004**, *45*, 573−579.

(11) Kubelka, J.; Hofrichter, J.; Eaton, W. A. *Curr. Opin. Struct. Biol.* **2004**, *14*, 76−88.

(12) Naganathan, A. N.; Munoz, V. *J. Am. Chem. Soc.* **2005**, *127*, 480−481.

(13) Klimov, D. K.; Thirumalai, D. *J. Comput. Chem.* **2002**, *23*, 161−165.

(14) Li, M. S.; Klimov, D. K.; Thirumalai, D. *Phys. Rev. Lett.* **2004**, *93*, 268107−268110.

(15) Li, M. S.; Klimov, D. K.; Thirumalai, D. *Physica A* **2005**, *350*, 38−44.

(16) Holtzer, M. E.; Loett, E. G.; d'Avignon, D. A.; Holtzer, A. *Biophys. J.* **1997**, *73*, 1031−1041.

(17) Ma, H. R.; Gruebele, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 2283−2287.

(18) Lakshmikanth, G. S.; Sridevi, K.; Krishnamoorthy, G.; Udgaonkar, J. B. *Nat. Struct. Biol.* **2001**, *8*, 799−804.

(19) Kohn, J. E.; Millett, I. S.; Jacob, J.; Zagrovic, B.; Dillon, T. M.; Cingel, N.; Dothager, R. S.; Seifert, S.; Thiyagarajan, P.; Sosnick, T. R.; Hasan, M. Z.; Pande, V. S.; Ruczinski, I.; Doniach, S.; Plaxco, K. W. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 12491−12496.

(20) Fisher, M. E.; Berker, A. N. *Phys. Rev. B* **1982**, *26*, 2507−2513.

(21) Honeycutt, J. D.; Thirumalai, D. *Biopolymers* **1992**, *32*, 695−709.

(22) Go, N. *Annu. Rev. Biophys. Bioeng.* **1983**, *12*, 183-210.

(23) Clementi, C.; Nymeyer, H.; Onuchic, J. *J. Mol. Biol.* **2000**, *298*, 937−953.

(24) Camacho, C. J.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 6369−6372.

(25) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1989**, *63*, 1195−1198.

(26) Jackson, S. E.; Fersht, A. R. *Biochemistry* **1991**, *30*, 10428−10435.

(27) Klimov, D. K.; Thirumalai, D. *Folding Des.* **1998**, *3*, 127−139.

(28) Kaya, H.; Chan, H. S. *Proteins: Struct., Funct., Genet.* **2000**, *40*, 637−661.

(29) Kaya, H.; Chan, H. S. *J. Mol. Biol.* **2003**, *326*, 911−931.

(30) Chan, H. S.; Shimizu, S.; Kaya, H. *Methods Enzymol.* **2004**, *380*, 350−379.

(31) Kaya, H.; Chan, H. S. *Phys. Rev. Lett.* **2000**, *85*, 4823−4826.

(32) Klimov, D. K.; Thirumalai, D. *J. Chem. Phys* **1998**, *109*, 4119−4125.

(33) Goldstein, R. A.; Lutheyschulten, Z. A.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 4918−4922.

(34) Cheung, M. S.; Garcia, A. E.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 685−690.

(35) Gutin, A. M.; Abkevich, V. I.; Shakhnovich, E. I. *Phys. Rev. Lett.* **1996**, *77*, 5433−5436.

(36) Li, M. S.; Klimov, D. K.; Thirumalai, D. *J. Phys. Chem. B* **2002**, *106*, 8302−8305.

(37) Koga, N.; Takada, S. *J. Mol. Biol.* **2001**, *313*, 171−180.

(38) Betancourt, M. R. *J. Chem. Phys.* **1998**, *109*, 1545−1554.

(39) A complete list of wild-type proteins, the values of cooperativity measure Ω_c, and the folding transition width $\Delta T/T_F$ with corresponding errors are available at www.biotheory.umd.edu/ScalingDB.html.

(40) Hyeon, C.; Thirumalai, D. *Biochemistry* **2005**, *44*, 4957−4970.

(41) Dyer, R. B. Unpublished results.

(42) Xu, Y.; Oyola, R.; Gai, F. *J. Am. Chem. Soc.* **2003**, *125*, 15388−15394.

(43) Honda, S.; Kobayashi, N.; Munekata, E. *J. Mol. Biol.* **2000**, *295*, 269−278.

(44) Qiu, L.; Pabit, S. A.; Roitberg, A. E.; Hagen, S. J. *J. Am. Chem. Sci*. **2002**, *124*, 12952−12953.

(45) Williams, S.; Causgrove, T. P.; Gilmanshin, R.; Fang, K. S.; Callender, R. H.; Woodruff, W. H.; Dyer, R. B. *Biochemistry* **1996**, *35*, 691−697.

(46) Kubelka, J.; Eaton, W. A.; Hofrichter, J. *J. Mol. Biol.* **2003**, *329*, 625−630.

(47) Ferguson, N.; Johnson, C. M.; Macias, M.; Oschkinat, H.; Fersht, A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 13002−13007.

(48) Ferguson, N. Private communication.

(49) Kouza, M.; Chang, C. F.; Hayryan, S.; Yu, T. H.; Li, M. S.; Huang, T. H.; Hu, C. K. *Biophys. J.* **2005**, *89*, 3353−3361.

(50) Alexander, P.; Fahnestock, S.; Lee, T.; Orban, J.; Bryan, P. *Biochemistry* **1991**, *31*, 3597−3603.

(51) Knapp, S.; Karshikoff, A.; Berndt, K. D.; Christova, P.; Atanasov, B.; Ladenstein, R. *J. Mol. Biol.* **1996**, *264*, 1132−1144.

(52) Wassenberg, D.; Welker, C.; Jaenicke, R. *J. Mol. Biol.* **1999**, *289*, 187−193.

(53) Knapp, S.; Mattson, P. T.; Christova, P.; Berndt, K. D.; Karshikoff, A.; Vihinen, M.; Smith, C. I. E.; Ladenstein, R. *Proteins: Struct., Funct., Genet.* **1998**, *31*, 309−319.

(54) Roy, S.; Hechts, M. H. *Biochemistry* **2000**, *39*, 4603−4607.

(55) Villegas, V.; Azuaga, A.; Catasus, L.; Reverter, D.; Mateo, P. L.; Aviles, F. X.; Serrano, L. *Biochemistry* **1995**, *34*, 15105−15110.

(56) Naik, M.; Huang, T.-h. *Protein Sci.* **2004**, *13*, 2483−2492.

(57) Clarke, J.; Hamill, S. J.; Johnson, C. M. *J. Mol. Biol.* **1997**, *270*, 771−778.

(58) Pace, C. N.; Hebert, E. J.; Shaw, K. L.; Schell, D.; Both, V.; Krajcikova, D.; Sevcik, J.; Wilson, K. S.; Dauter, Z.; Hartley, R. W.; Grimsley, G. R. *J. Mol. Biol.* **1998**, *279*, 271−286.

(59) Van Nuland, N. A. J.; Meijberg, W.; Warner, J.; Forge, V.; Ruud, M. Scheek, R. M.; Robillard, G. T.; Dobson, C. M. *Biochemistry* **1998**, *37*, 622−637.

(60) Martinez, J. C.; Elharrous, M.; Filimonov, V. V.; Mateo, P. L.; Fersht, A. R. *Biochemistry* **1994**, *33*, 3919−3926.

(61) Arnold, U.; Ulbrich-Hofmann, R. *Biochemistry* **1997**, *36*, 2166−2172.

(62) Hirai, M.; Arai, S.; Iwase, H. *J. Phys. Chem.* **1999**, *103*, 549.

(63) Makhatadze, G.; Clore, G. M.; Gronenborn, A. M.; Privalov, P. L. *Biochemistry* **1994**, *33*, 9327−9332.

(64) Munoz, V. M.; Eaton, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11311−11326.

(65) Galzitskaya, O. V.; Finkelstein, A. V. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11299−11304.

(66) Alm, E.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11305−11310.