Three-body interactions improve the prediction of rate and mechanism in protein folding models

M. R. Ejtehadi[†‡§], S. P. Avall[†], and S. S. Plotkin[†§]

[†]Department of Physics and Astronomy, University of British Columbia, Vancouver, BC, Canada V6T 1Z1; and [‡]Department of Physics, Sharif University of Technology, Tehran 11365-9161, Iran

Edited by Peter G. Wolynes, University of California at San Diego, La Jolla, CA, and approved August 19, 2004 (received for review May 18, 2004)

Here we study the effects of many-body interactions on rate and mechanism in protein folding by using the results of molecular dynamics simulations on numerous coarse-grained C^{α}-model single-domain proteins. After adding three-body interactions explicitly as a perturbation to a Gō-like Hamiltonian with native pairwise interactions only, we have found (*i*) a significantly increased correlation with experimental ϕ values and folding rates, (*ii*) a stronger correlation of folding rate with contact order, matching the experimental range in rates when the fraction of three-body energy in the native state is ≈20%, and (*iii*) a considerably larger amount of three-body energy present in chymotripsin inhibitor than in the other proteins studied.

Understanding the nature of the interactions that stabilize protein structures and govern protein folding mechanisms is a fundamental problem in molecular biology (1–6) that has applies to structure and function prediction (7–10) as well as rational enzyme design (11). Regarding folding mechanisms, protein folding has long been known to be a cooperative process, at least for smaller single-domain proteins (12). Experimental scenarios that lack a first-order-like folding barrier are rare (13), often in contrast to simulation results. There are other discrepancies between simulation and experiment. For example, although the experimental folding rates for a typical set of 18 two-state, single-domain proteins (given in *Materials and Methods*) span about six orders of magnitude, simulations of coarse-grained models of the same proteins have rates that vary by about a factor of 100, a discrepancy of four orders of magnitude.

How does one then quantify the sources of the barrier that controls the folding rate? The folding barrier is the residual of an incomplete cancellation of large and opposing energetic and entropic contributions, with the relative smallness of the barrier allowing folding to occur on biological time scales (14, 15). Among the important energetic contributions that drive folding are solvent-mediated hydrophobic forces (16), which are known to be weaker on short-length scales, or low concentrations of apolar side-chains (17), a scenario likely to be present when the protein is unfolded. Hence, the solvent-averaged potential governing folding almost certainly contains a nonadditive, many-body component, and several models have been proposed to capture this effect (18–27). The folding free-energy barrier increases as the nonadditivity of interactions is increased (20, 21, 23, 25) because of the decreased energetic correlation between the native conformation and conformations that may be geometrically similar to it.

Experimental ϕ values give a measure of the strength of native interactions involving a particular amino acid (residue) in the transition state (28), thus quantifying a residue's importance in folding. However the ϕ values obtained from simulations of coarse-grained protein models generally do not correlate well with the experimentally determined values. Model proteins are coarse-grained based on the belief that a reduced number of degrees of freedom can capture the essentials of the folding process (4, 29, 30); however, the less than ideal agreement with experimentally observed rates and mechanisms leads one to consider alternate forms for the coarse-grained Hamiltonian or energy function as well as more detailed all-atom models (31–33) that may contain explicit solvent as well (6, 33–38).

But it is also clear that coarse-grained simulations allow a study of microscopic dynamics that would not be possible by all-atom models with present-day computing power. Because we cannot yet fully analyze the statistics of folding trajectories in all-atom models, coarse-grained simulational models, such as off-lattice C^{α} models (4, 30, 39–43) have been essential in elucidating protein-folding mechanisms.

We could then take the following approach: postulate a given feature thought to be present in the system and ask to what extent this feature, such as many-body potentials, must be present in the Hamiltonian of a coarse-grained model for best agreement with existing experimental data on protein folding rates and mechanisms.

Materials and Methods

Simulation Model. Eighteen two-state folding proteins with known native structures [Protein Data Bank (PDB) ID codes 1AEY, 1APS, 1FKB, 1HRC, 1MJC, 1NYF, 1SRL, 1UBQ, 1YCC, 2AIT, 2CI2, 1PTL, 2U1A, 1AB7, 1CSP, 1LMB, 1NMG, 1SHG] were selected for coarse-grained simulations. For all proteins except the last five above, rate data were available at various denaturant concentrations. These proteins were then used for further analysis at the stability of the transition midpoint.

The simulated proteins consist of a chain of connected beads, with each bead representing the position of the C^{α} atom in the corresponding amino acid. The off-lattice C^{α} Gō model has been described in detail in refs. 30, 39, 43, and 44. The Hamiltonian has local and nonlocal parts: Bond, angle, and dihedral angle potentials constitute local interactions. In the putative Gō model, pair contacts between residues in spatial proximity in the native structure constitute nonlocal interactions. Nonnative interactions are treated by a sterically repulsive pair-potential only.

Heavy atoms within a cutoff distance of $r_c = 4.8$ Å in the native structure obtained from the PDB file are associated with a Lennard–Jones-like 10–12 potential of depth $\varepsilon_2 = -k_B T$ and a position of the minimum equal to the distance of the C^{α} atoms in the native structure. Let there be N_2 pair contacts of energy ε_2 in the native PDB structure. Then in an arbitrary conformation there are QN_2 contacts with energy $E_2 \approx \varepsilon_2 QN_2$, with Q being the fraction of native pair contacts (we account for the continuum nature of the Lennard–Jones potentials).

We let triples with heavy atoms within a cutoff distance of 4.8 Å in the native structure have an energy ε_3. For a given protein there will then be N_3 three-body contacts present in the PDB native structure, with total three-body energy $\varepsilon_3 N_3$. An arbitrary structure then has a three-body contribution to the energy of $E_3 \equiv \varepsilon_3 Q_3 N_3$,

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: rmsd, rms distance; TTSE, thermal transition state ensemble; KTSE, kinetic transition state ensemble; CI2, chymotripsin inhibitor 2; SH3, src homology 3; FKBP, FK506-binding protein; AcP, acylphosphatase; MJ, Miyazawa–Jernigan; PDB, Protein Data Bank.

[§]To whom correspondence may be addressed. E-mail: steve@physics.ubc.ca or ejtehadi@physics.ubc.ca.

www.pnas.org/cgi/doi/10.1073/pnas.0403486101

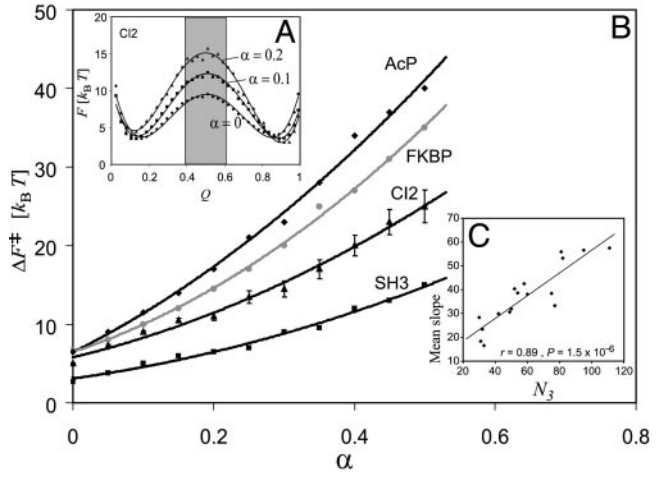

Fig. 1. The folding barrier height $\Delta F^{\neq}$ increases with increasing three-body contribution to the energy α. (*A*) Free energy versus the fraction of native contacts Q for CI2 for three values of α. (*B*) The barrier versus α for four proteins selected from Table 1. Shown for CI2 are error bars obtained from the standard deviation of F(Q) by using a bin size $\Delta Q = 4/149$. (*C*) The average slope of $\Delta F^{\neq}$ versus α correlates strongly with the number of three-body interactions in the native state ($r = 0.89$, $P = 10^{-6}$). Therefore, the barriers in *B* increase at different rates because of differing numbers of triples formed in the transition states of the various proteins: More native triples typically means a larger three-body contribution to the barrier. The shaded region in *A* corresponds to the TTSE described in *Materials and Methods*. In general, this ensemble depends on α.

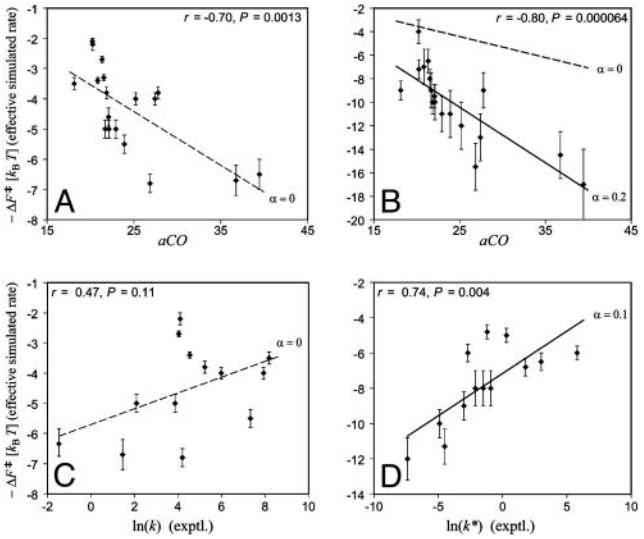

Fig. 2. Comparison of simulated and experimental rates. (*A*) Simulated folding barriers (effectively measuring logarithm folding rates for 18 proteins listed in *Materials and Methods*) for a pairwise interacting Gō model correlate well with absolute contact order (aCO) (43). (*B*) Simulated folding barriers show an increased correlation with absolute contact order when the fraction of native three-body energy is such that the dispersion in effective simulated rates matches the experimental dispersion for this dataset ($\alpha = 20\%$). Rates now span 5.7 decades, in contrast to 2 decades for a pure two-body Hamiltonian (dashed line in *B* is the best fit line in *A*). (*C*) For 13 of the 18 proteins (see *Materials and Methods* for a list), rate data were available for various different denaturant concentrations. These proteins were used for the analysis in *C* and *D*. For these proteins, the simulated effective log rates do not correlate significantly with the experimental rate data at 25°C. (*D*) By tuning the rate data to the transition midpoints and introducing three-body energy in the native state, we saw a significant increase in the correlation between experimental and simulated rate data, with best correlation when $\alpha = 10\%$.

where Q_3 is the fraction of native triples present in that conformation. Three-body interactions are again Gō-like; the remaining bond, angle, dihedral, and nonnative interaction energies are all unchanged.

When both pairwise and three-body interactions are present, the native nonlocal part of the energy becomes

$$E_{\mathrm{NL}}(\alpha) = (1 - \alpha)E_2 + \alpha E_3. \qquad [1]$$

The free parameter α ($0 \leq \alpha \leq 1$) controls the relative contribution of two- and three-body interactions. The energy per triple is assigned as $\varepsilon_3 = \varepsilon_2 N_2/N_3$ to preserve overall native stability.

Dense sampling is obtained from long simulations with a purely two-body Gō Hamiltonian at the transition midpoint [e.g., for chymotripsin inhibitor 2 (CI2) the simulation time corresponds to ≈3 sec, as determined from the number of folding and unfolding events). From histograms of the number of states at a given fraction of native contacts Q, the free energy $F(Q)$ can be constructed. All simulated free energy profiles displayed a single dominant barrier. All proteins are considered at their transition midpoints only, when the unfolded and folded free energies are equal: $F_{\mathrm{U}} = F_{\mathrm{F}}$ (Fig. 1*A*).

Three-body energies are treated as a perturbation on the Hamiltonian. The new free energy is given by the exact expression

$$\frac{F(Q,\alpha)}{k_{\mathrm{B}}T} = -\ln \frac{\sum_i \mathrm{e}^{-\Delta E_i(\alpha)/k_{\mathrm{B}}T}\, \Delta(Q^{(i)},Q)}{\sum_i \mathrm{e}^{-\Delta E_i(\alpha)/k_{\mathrm{B}}T}}, \qquad [2]$$

where the sum is on all sampled conformations i, $\Delta(Q^{(i)},Q)$ is a delta function that selects only those states where $Q^{(i)} = Q$, and $\Delta E(\alpha) = E_{\mathrm{NL}}(\alpha) - E_2$. Fluctuations in F($Q$,$\alpha$) arise from both finite sampling and the fact that configurations with similar Q may have different numbers of three-body interactions. We found that the latter inherent effect dominated the fluctuations; however, the free energy barriers were still well determined after binning over small ranges of Q ($\Delta Q \approx 0.02$, see the error bars in Fig. 1*A*).

Calculated ϕ Values. Simulated kinetic ϕ values (45) are given by

$$\phi_i = \frac{\langle n_i \rangle_{\neq} - \langle n_i \rangle_U}{\langle n_i \rangle_F - \langle n_i \rangle_U}, \qquad [3]$$

where $\langle n_i \rangle$ is the thermal mean value of the number of contacts for residue i, and the $\neq$, U, and F subscripts refer to the transition state, unfolded state, and folded state ensembles, respectively.

We first compared simulated and experimental ϕ values by using the thermal transition state ensemble (TTSE) around the free energy barrier peak, i.e., $|F\text{-}F^{\neq}|/\Delta F^{\neq} \leq 0.2$ was used to define a width ΔQ of the barrier peak (Fig. 1*A*, shading). Conformations within this range were taken to be the TTSE and were used to calculate ϕ values from Eq. **3**. The validity of the TTSE was checked for CI2 and src homology 3 (SH3) with a comparison of ϕ values by using the kinetic transition state ensemble (KTSE), selected as having a folding probability p_{FOLD} of roughly 1/2 (46). Conformations in the TTSE were used as initial conditions for 100 simulations that were terminated when the protein folded or unfolded. Those conformations that had a p_{FOLD} within $0.5 \pm 1/\sqrt{100}$ were taken as the KTSE. For CI2 (SH3) we found 315 (283) KTSE configurations from a total of 2,359 (2,078) TTSE configurations.

Other reaction coordinates were helpful in determining the KTSE by constructing multidimensional reaction surfaces. To this end we found a contact-order-weighted variant of Q to be useful, which for any configuration ν is given by

$$Q_{\mathrm{co}}^{\nu} = \frac{\sum_{i<j} |i-j|\, \Delta_{ij}^{\nu}\, \Delta_{ij}^{N}}{\sum_{i<j} |i-j|\, \Delta_{ij}^{N}}, \qquad [4]$$

where the sum is over all C^α atoms, and Δ_{ij}^{ν} and Δ_{ij}^{N} are unity if residues i and j are in contact in conformations ν and the native structure respectively, otherwise they are zero.

We determined ϕ values in the presence of three-body interactions analogously to Eq. **3**. Under some simplifying assumptions (e.g., requiring a ϕ value that is independent of the perturbation energies),

$$\phi_i^{(\alpha)} = \frac{(1-\alpha)(\langle n_i\rangle_{\neq}^{(\alpha)} - \langle n_i\rangle_U^{(\alpha)})N_3 + \alpha(\langle m_i\rangle_{\neq}^{(\alpha)} - \langle m_i\rangle_U^{(\alpha)})N_2}{(1-\alpha)(\langle n_i\rangle_F^{(\alpha)} - \langle n_i\rangle_U^{(\alpha)})N_3 + \alpha(\langle m_i\rangle_F^{(\alpha)} - \langle m_i\rangle_U^{(\alpha)})N_2}. \quad [5]$$

Here, m_i is the number of three-body interactions in which monomer i is involved, and superscript (α) indicates averaging the ensembles ($\neq$, U, and F) in the presence of three-body energy. When $\alpha \to 0$, Eq. **5** reduces to Eq. **3**.

Miyazawa–Jernigan (MJ)-Based Models. The effect of heterogenity in the model was also studied by interpolating between the Gō model and the MJ models by varying the free parameter α between zero (homogeneous Gō model) and unity (MJ model). The contact energy for any pair of residues (not necessarily native) is then

$$\varepsilon_{ij} = (1-\alpha)\varepsilon_2 + \alpha\varepsilon_{ij}^{\mathrm{MJ}}, \quad [6]$$

where ε_2 is as above and $\varepsilon_{ij}^{\mathrm{MJ}}$ is proportional to the MJ interaction energy (47) between the residue types of i and j, scaled by a factor to ensure that the energy of the native structure is α-independent. An interpolation between a uniform Gō model and a heterogeneous Gō model with native contact energies given by MJ parameters was also considered.

Contact Order and Statistical Significance. Absolute contact order is the average sequence separation between residues having native contacts (48): $aCO = M^{-1}\Sigma_{i>j}|i-j|$, where M is the total number of native contacts. Relative contact order is scaled again by chain length N: $rCO = aCO/N$.

Statistical significance or P value is the probability to achieve a given correlation coefficient, r, assuming random data: $P = \mathrm{erf}(|r|\sqrt{N/2})$. Small datasets almost always have fairly large P values, even if r is large. Large datasets may still have small P values even if the correlation is weak, which would still indicate a systematic effect.

Results

Protein Folding Rates. Here we considered the effect of introducing a three-body potential to an off-lattice two-body Gō model studied in refs. 43, 44, and 49. Eighteen above-mentioned single-domain proteins that are known to fold by a two-state mechanism were selected and coarse-grained so that each amino acid corresponded to a bead at the position of the C^α atom. Long simulations at the folding temperature T_f for a subset of the proteins showed a single exponential distribution of first passage times: $P(\tau) \sim \exp(-\kappa t)$. For these proteins, the simulated log folding rate, $\log(\kappa)$, correlated very strongly ($r = 0.997$) with the free energy barrier height $\Delta F^{\neq}$, indicating that $\Delta F^{\neq}$ was an accurate predictor of the rate for the simulated Gō models. We subsequently assume this proportionality between $\Delta F^{\neq}$ and $-\log(\kappa)$ for all simulated proteins, referring to $\exp(-\Delta F^{\neq}/k_BT)$ as the "effective rate."

The above mentioned discrepancy between the effective protein rates for our dataset and the experimentally determined rates for the same proteins motivates an investigation of the effect of many-body interactions on rates. When a portion of the total energy is attributable to many-body interactions, energetic gain is not achieved until a larger amount of native structure is present, with a correspondingly larger entropic cost. Several polymer loops must be simultaneously closed during folding to receive energetic gain. This effect enhances the dependence of rate on contact order, increasing the range over which rates vary.

By attributing a fraction α of the native energy to triples in the native structure, we studied the effects of three-body interactions by varying this single parameter (see *Materials and Methods*). The effects on the free energetic potential surface for several proteins are shown in Fig. 1*B*.

As the fraction of three-body energy is increased, the correlation of the simulated effective rates with both absolute and relative contact order and the range of values over which rates vary increase (Fig. 2 *A* and *B*). Similar effects have also been seen in lattice protein models (50, 51). We can also quantify how much three-body energy at the residue level reproduces the experimental dispersion in rates for single-domain proteins. The simulated effective rates span six orders of magnitude when ≈20% of the energy in the native state of the coarse-grained protein is due to three-body interactions.

Rates simulated with a two-body Hamiltonian do not correlate significantly with experimentally determined rates at 25°C (Fig. 2*C*). We can remove the effects due to variations in stability and reflect the conditions in the simulations by taking instead the rate data at the various transition midpoints (after the addition of GdHCl). We then found the correlation significantly increased to $r = 0.64$ and $P = 0.018$. Adding three body energy in the simulations increases the correlation with the experimental rates (at the transition midpoints) still further, with the best correlation achieved when $\alpha = 10\%$ (see Fig. 2*D*).

These results strongly suggest that (*i*) stability is an important determinant of folding rate, (*ii*) many-body energy is present in the energy functions of real proteins, and (*iii*) Gō or Gō-like models (which ignore nonnative interactions) can predict experimental rates, illustrating the minor importance of nonnative interactions in governing folding barriers.

The correlation of log rates with *rCO* also improves as α is increased from zero; however, the correlations are modest, increasing from $r = -0.29$ and $P = 0.24$ at $\alpha = 0$ to a best correlation of $r = -0.44$ and $P = 0.08$ at $\alpha = 10\%$ (data not shown).

Testing Pair-Interaction Matrices. The correlation between experimental and simulational ϕ values for a two-body Hamiltonian (r_0, P_0) was typically not statistically significant (see Table 1), with the exception of SH3. Rank-ordered measures of correlation, such as Kendall's τ, which are insensitive to the precise values of the data, generally do not improve the agreement (Table 2). We also checked whether simulations with a two-body Hamiltonian could accurately predict residues that had higher ϕ values. This calculation was done by weighting the statistical averaging in the correlation coefficient by the experimental ϕ value itself as a Jacobian factor. Implementing this recipe did not substantially increase the correlation coefficient and, in fact, decreased it in the cases of acylphosphatase (AcP) and CI2 (Table 1). Similar results were obtained by implementing a simple cutoff imposing a lower bound for relevant experimental ϕ values (data not shown).

The experimental data can be used to test energy functions characterizing pair interactions at the amino acid level, such as the MJ matrix (47). We investigated whether MJ interaction parameters improved the simulational predictions of ϕ values by interpolating between a homogeneous Gō model and a model with pair interactions (between all residues) governed by MJ parameters (see Eq. **5**). We also interpolated between a homogeneous Gō model and a heterogeneous Gō model with native interaction parameters determined from the MJ matrix.

Results are shown for two proteins in Fig. 3. For CI2 and SH3, no improvements in the correlation with experimental data were seen by implementing this procedure. Table 1 shows the results for the comparison between experimental ϕ-value data and ϕ values obtained from a pairwise MJ Hamiltonian. In general, if correlations increased by interpolating toward MJ parameters, they did so only modestly: Only in the case of protein L did the improvement reach statistical significance ($P = 1\%$, see Table 1).

To check of the validity of the recipe of interpolating toward MJ

Table 1. Two-body and three-body characterization of proteins studied

Models	Proteins				
	SH3	FKBP	AcP	Protein L	CI2
Gō					
r_0	0.58†	0.32	0.12	0.18	−0.10†
P_0	0.0003	0.17	0.58	0.25	0.56‡
MJ					
α^*, %	0	10	50	20	0
r_{α^*}	0.59	0.41	0.35	0.38	−0.017
P_{α^*}	0.0003	0.07	0.1	0.01	0.92‡
MJ-Gō					
α^*, %	5	20	30	30	0
r_{α^*}	0.59	0.38	0.30	0.38	−0.017
P_{α^*}	0.0002	0.1	0.16	0.01	0.92‡
Three-body					
α^*, %	5	10	15	15	35
r_{α^*}	0.60†	0.43	0.32	0.53	0.57†
P_{α^*}	0.0001	0.057	0.14	0.00027	0.0004
N	56	107	98	62	65
N_2	128	299	257	126	148
N_3	32	111	97	30	54
n	35	20	23	41	35
$\Delta F^{\neq}_{\alpha^*}$	3.8 ± 0.2	10 ± 0.8	14 ± 2.0	6.2 ± 0.5	17 ± 3.5
$\Delta F^{\neq}_{\alpha^*}/\Delta F^{\neq}_0$	1.4	1.5	2.2	2.8	3.4
$E^{\neq}_{3B}/E^{\neq}_{tot}$, %	2.6†	5.5	8.9	3.3	13.0†
High ϕ					
$\tilde{r}_0$	0.65†	0.37	−0.02	0.26	−0.43†
$\tilde{P}_0$	2.7×10^{-5}	0.10	0.91‡	0.10	0.01‡

The sources for experimental ϕ-value data for SH3, FKBP, AcP, CI2, and protein L (PDB ID codes 1SRL, 1FKB, 1APS, 2CI2, and 2PTL, respectively) are refs. 54, 57, 56, 58, and 59, respectively. The Gō model data comprises the correlation coefficient and statistical significance between experiments and simulation of a pairwise interacting Gō model. α^* is in general the value of the interpolation parameter that gives best agreement with the experimental data for each corresponding model. For the MJ models, Eq. **6** is used; for the three-body models, Eq. **1** is used. r_{α^*} and P_{α^*} are the correlation coefficient and statistical significance, respectively, at best agreement for each corresponding model. N is chain length; N_2 is the number of native pair contacts; N_3 is the number of native triples; n is the number of ϕ-value data points used in the comparison; $\Delta F^{\neq}_{\alpha^*}$ is the barrier height in k_BT at α^* for the three-body model; $\Delta F^{\neq}_{\alpha^*}/\Delta F^{\neq}_0$ is the ratio of the free energy barriers when $\alpha = \alpha^*$ and $\alpha = 0$; and $E^{\neq}_{3B}/E^{\neq}_{tot}$ *is the fraction of three-body energy in the transition state ensemble at* α^*. *For high-*ϕ *weighting,* $\tilde{r}_0$ and $\tilde{P}_0$ are the correlation coefficient and statistical significance, respectively, including a Jacobian factor weighting each term in the correlation function by the experimental ϕ value itself, i.e. averages are calculated as $\langle A \rangle = (\Sigma_1^n \phi_i^{exp} A_i)/(\Sigma_1^n \phi_i^{exp})$, where n is the number of data points. This recipe simply stresses the importance of the agreement between large ϕ values.

†KTSE was used.

‡We allow for the possibility of anticooperativity in proteins and, hence, ascribe statistical significance to negative correlations. Thus, P values here are two-sided.

Table 2. Kendall's τ and statistical significance between experiment and simulation

Models	Proteins				
	SH3	FKBP	Protein L	AcP	CI2
Gō					
τ_0	0.42†	0.27	0.14	0.14	0.042†
P_0	0.00044	0.10	0.19	0.37	0.72
Three-body					
α^*, %	0	10	20	25	35
τ_{α^*}	0.42†	0.31	0.36	0.33	0.40†
P_{α^*}	0.00044	0.055	0.00069	0.027	0.0008

Kendall's τ measure of ranked correlation and statistical significance [$P(|\tau'| \geq |\tau|)$] of τ value between experiments and simulations for a pairwise interacting Gō model and the two- plus three-body model. α^* is the value of the interpolation parameter that gives best agreement with experimental data for a two- plus three-body Hamiltonian as in Eq. **1**.

†KTSE was used.

parameters, we compared the largest improvement in correlation ($r_{\alpha^*} - r_o$) with the value α^* of MJ energy in Eq. **6** required to achieve that correlation. This test determines whether the poorness of the original correlation was due to the absence of MJ coupling energies. We found that ($r_{\alpha^*} - r_o$) itself correlated well with α^*; however, the statistical significance was not particularly strong, and the slope measuring the degree of improvement was not particularly high (Fig. 4).

Testing Three-Body Interactions. The experimental data can also be used as a benchmark to test what amount of three-body energy in the Hamiltonian of the coarse-grained model gives best agreement with experimental ϕ values. We examined this question for the five proteins listed in Table 1 by measuring the correlation between the experimentally obtained ϕ values and ϕ values of the same residues determined from simulations, with conditions ranging from between a pairwise interacting Gō model protein and one governed exclusively by three-body interactions at the residue level (see *Materials and Methods*).

As the strength of three-body interactions increased from zero, the correlation coefficient also increased for all proteins studied (Fig. 3 and Table 1). An exceptional case was SH3, which showed only a modest increase in correlation for the KTSE and no increase for the TTSE. The fraction α^* of native three-body energy that gave best agreement with experimental data varied from protein to protein but correlated strongly with the increase in agreement with experimental data (see Table 1). That is, the improvement in correlation ($r_{\alpha^*} - r_o$) itself correlated very strongly with α^* ($r = 0.97, P = 0.005$), further supporting the notion that the poorness of the original agreement was due at least in part to the absence of many-body forces (see Fig. 4).

For a protein with a large fraction of three-body energy, such as CI2, the transition states in the presence of three-body interactions is significantly different from the two-body transition state. For CI2, the rms distance (rmsd) between all 315 structures in the KTSE was found for both the two-body and two- plus three-body (at α^*) cases. From the rmsd, the "most representative" transition state structure may be defined as having the minimal Boltzmann-weighted rmsd (minimum over structure i of $\Sigma_j p_j(\text{rmsd})_{ij}$) to all others in the KTSE. The two-body case shows more overall secondary structure, in particular more α-helix but less β-sheet. The Q, Q_{CO} (see *Materials and Methods*), and R (rmsd from the native structure) values are shown in Fig. 6, which is published as supporting information on the PNAS web site. These findings indicate that the two- plus three-body transition state is less structured than the pure two-body transition state. However, kinetically the structures are about the same distance from the native in that their p_{FOLD} values are comparable (see Fig. 6). The structures have a rmsd of 7.8 Å between them, so they are structurally distinct from each other. Interestingly, the high-ϕ residue 34 has more local secondary structure in the pure two-body case than at α^*; it also has no triples in the native state and its high ϕ value in the presence of three-body interactions is the result of correlations with other triples made in the transition state.

The procedure of adding three-body interactions was repeated considering only residues in the hydrophobic core of native structure, in this case buried with less than ≈30% accessible surface area, by using the Swiss–PDB Viewer (www.expasy.org/spdbv). We saw qualitatively the same effect, but the change in correlation coefficient was less pronounced, increasing to ≈0.42 for CI2, for example.

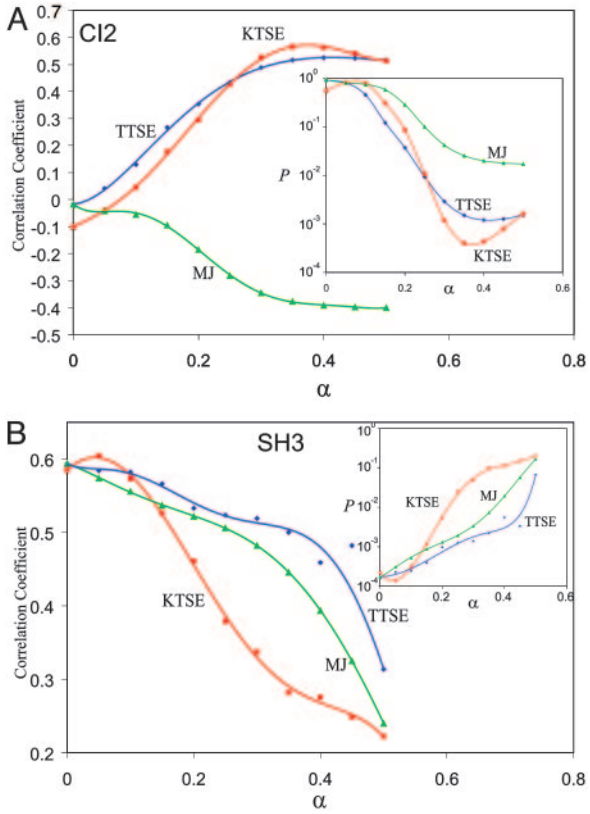

Fig. 3. Comparison of the agreement of ϕ values between simulation and experiment for CI2 (*A*) and SH3 (*B*). Green curves show the correlation coefficient and statistical significance (*Insets*) for ϕ values derived from the TTSE in the simulations as the Hamiltonian was continuously changed from a uniform Gō model to one with pair interactions governed by MJ parameters (the curve shown in *A Inset* is the statistical significance of the anticorrelation) (see Eq. **6**). No improvement was seen for CI2 or SH3 by implementing this recipe. Red and blue curves show the correlation coefficient and statistical significance between experimental and simulated ϕ values as a function of the fraction α of three-body energy in the native state. Blue curves correspond to TTSE; red curves correspond to KTSE. For CI2, the improvement as α is increased is dramatic, with best agreement with the experiment at ≈35% three-body energy. On the other hand, SH3 was exceptional in that it showed the opposite trend, with best agreement for a purely pairwise interacting model for the TTSE and $\alpha = 5\%$ for the KTSE. All other proteins studied were bracketed by these two extremes: They showed moderate components of three-body energy, with moderate to large increases in correlation coefficient (Table 1).

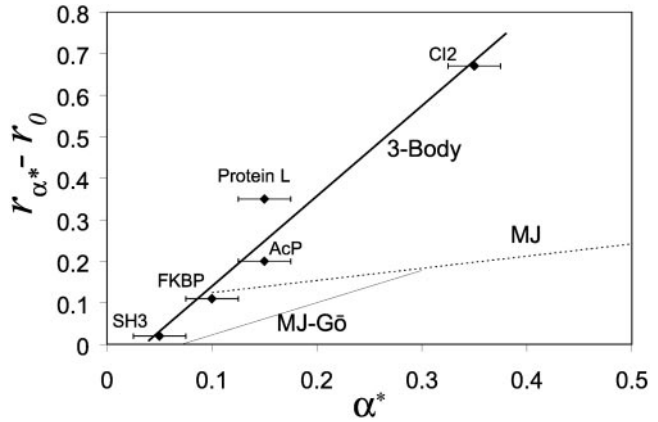

Fig. 4. Plot of the largest improvement in correlation ($r_{\alpha^*} - r_o$) vs. the value of interpolation parameter α^* required to achieve that correlation. Energy functions are interpolated toward a three-body Gō model (Eq. **1**) and two-body models with MJ energetic parameters (Eq. **6**). The slope and correlation indicate the validity of the interpolation procedure. Adding three-body energies gives a slope of 2.2, and ($r = 0.97$ and $P = 0.005$). Adding a MJ component to the pair interaction energies gives a slope of 0.29 but a fit that is not statistically significant ($r = 0.83$ and $P = 0.38$). Restricting the MJ component to native interaction energies gives a statistically significant fit ($r = 0.956$ and $P = 0.044$) but with a shallow slope (0.78), indicating only moderate improvement.

This finding implies that coarse-grained model proteins with effective solvent-averaged interactions have many-body interactions involving residues on the surface as well.

For further information, see *Supporting Text*, which is published as supporting information on the PNAS web site.

Discussion

The above results suggest that many-body interactions can play a significant role in governing the folding mechanisms of two-state proteins when described at the residue level. This conclusion seems quite evident upon comparing the statistical significance rows in Table 1 or Table 2 for the pure two-body Hamiltonian and the two-plus three-body Hamiltonian at α^*. In essentially all cases, many-body interactions helped to establish consistency with protein folding experiments. Some proteins showed dramatic improvement and others showed mild improvement, so proteins may be additionally classified through this effect. The value of α^* may be used as an indication of the importance of many-body interactions in governing the folding mechanism for a given protein, as the proteins are ranked in Tables 1 and 2, for example.

Experimental rates vary by about four orders of magnitude more than rates obtained from coarse-grained models with two-body Hamiltonians. However, a modest three-body component to native stability (≈20% on average) was sufficient to reproduce the experimental variability in folding rates. Similar numbers for the three-body energy have been obtained from triple-mutant studies of barnase (52). It is an open question as to how large the many-body component might be in finer-scale and all-atom models of proteins. Quantifying this component in terms of the missing degrees of freedom of either protein or solvent is nontrivial. Even all-atom, explicit-solvent models may have large many-body effects:

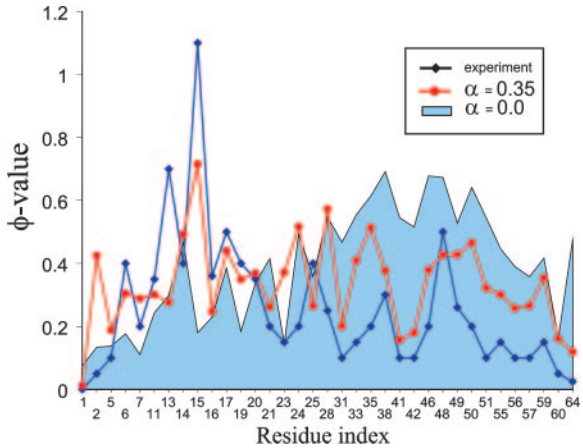

Fig. 5. ϕ value versus residue index for CI2, for experiment (blue trace), simulated pairwise Gō model (light-blue background), and two- plus three-body Gō model (red trace). The average ϕ values for the various energy functions are $\bar{\phi}^{(\mathrm{Expt})} = 0.25$, $\bar{\phi}^{(2)} = 0.40$, $\bar{\phi}^{(2+3)} = 0.33$, again confirming the more accurate two- plus three-body transition state is less structured. It is worth noting that native state is more stable in the experiments than in the simulations: The native stability is fixed at the transition midpoint in the simulations, regardless of the value of α.

Ab initio studies of interaction energies and reconfiguration barriers in water clusters suggest many-body energies can be quite significant (53).

For FK506-binding protein (FKBP), protein L, and CI2, the correlation between experimental and simulational ϕ values goes from insignificant to significant as three-body interactions are added. In the case of CI2, the agreement between simulations with a two-body energy function and experimental data were the poorest of the proteins studied, the fraction of three-body energy at best agreement was the largest, and the improvement in correlation coefficient was the most dramatic. In the case of SH3 on the other hand, the folding mechanism appears to be governed more by topology than by energetic considerations. In some sense, this is an exception that proves the rule, as previous evidence supported a folding mechanism dominated by topological considerations (54, 55).

Interestingly, muscle AcP had the poorest improvement in mechanism prediction by adding three-body interactions, as measured by the correlation coefficient; its original ϕ-correlation for a two-body Gō model was the second poorest after CI2. AcP also required the largest amount of MJ interactions for best agreement with experimental ϕ values but still correlated poorly even at best agreement. Intriguingly, AcP is also the slowest known two-state folder at present yet a good two-state folder with no intermediates (56). The slow folding is likely due to large contact order, however, and it would be interesting in the future to apply the three-body recipe to a topologically similar but faster folding protein, such as human procarboxypeptidase A2. On the other hand, the improvement for AcP as measured by Kendall's τ does, in fact, become statistically significant and suggests a large three-body component. We are inclined to take this more robust measure of statistical significance more seriously. The discrepancy of r and τ indicates some large outliers in ϕ values, likely because of variations in native stabilizing interactions, which may exist for functional reasons. These fluctuations in native interaction strength are not captured by the uniform Gō model and two- plus three-body models.

The largest improvement in correlation ($r_{\alpha^*} - r_o$) with the value of interpolation parameter α^* required to achieve that correlation was used as a measure to test the validity of the three-body and MJ interpolation recipes. The results for the three-body interpolation recipe showed a strong statistically significant correlation with a large slope indicating large rate of improvement. The results for the heterogeneous MJ Gō model also showed improvement, however with smaller slope and smaller statistical significance. It is noteworthy that for the case of CI2, in which the three-body recipe does the best, the MJ recipe failed to improve the agreement with experiment.

For CI2, the transition state in the presence of three-body interactions shows less overall native structure than the purely two-body transition state, despite the better agreement with experimental ϕ values for the three-body case. However it is not clear whether this will be a general rule. In both cases, the transition state consists largely of a disordered form of the native topology, sufficiently disordered to be kinetically balanced between the folded and unfolded states.

The low levels of agreement between experiment and simulation for two-body Hamiltonians told a somewhat cautionary tale. Although a large body of evidence leaves little doubt as to the importance of native topology in governing folding mechanism, these results should serve to show that realistic aspects of the energy function, such as many-body component to native stability, should not be ignored.

We thank Cecilia Clementi and Baris Oztop for helpful discussions. S.S.P. acknowledges support from the Natural Sciences and Engineering Research Council and the Canada Research Chairs program.

1. Wolynes, P. G., Onuchic, J. N. & Thirumalai, D. (1995) *Science* **267,** 1619–1620.
2. Dobson, C. M., Sali, A. & Karplus, M. (1998) *Angew. Chem. Int. Ed.* **37,** 868–893.
3. Fersht, A. R. (1999) *Structure and Mechanism in Protein Science* (Freeman, New York), 1st Ed.
4. Mirny, L. & Shakhnovich, E. (2001) *Annu. Rev. Biophys. Biomol. Struct.* **30,** 361–396.
5. Dill, K. A. & Chan, H. S. (1997) *Nat. Struct. Biol.* **4,** 10–19.
6. Daggett, V. & Fersht, A. R. (2003) *Nat. Rev. Mol. Cell Biol.* **4,** 497–502.
7. Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. (1999) *Nature* **402,** 83–86.
8. Hao, M.-H. & Scheraga, H. (1999) *Curr. Opin. Struct. Biol.* **9,** 184–188.
9. Bonneau, R. & Baker, D. (2001) *Annu. Rev. Biophys. Biomol. Struct.* **30,** 173–189.
10. Vendruscolo, M. & Domany, E. (1998) *J. Chem. Phys.* **109,** 11101–11108.
11. Bolon, D. N., Voigt, C. A. & Mayo, S. L. (2002) *Curr. Opin. Struct. Biol.* **6,** 125–129.
12. Jackson, Sophie E. (1998) *Folding Des.* **3,** R81–R91.
13. Gruebele, M. (1999) *Annu. Rev. Phys. Chem.* **50,** 485–516.
14. Hao, M.-H. & Scheraga, H. A. (1994) *J. Phys. Chem.* **98,** 4940–4948.
15. Plotkin, S. S. & Onuchic, J. N. (2002) *Q. Rev. Biophys.* **35,** 111–167, 205–286.
16. Dill, K. A. (1990) *Biochemistry* **29,** 7133–7155.
17. Lum, K., Chandler, D. & Weeks, J. D. (1999) *J. Phys. Chem.* **103,** 4570–4577.
18. Kolinski, A., Godzik, A. & Skolnick, J. (1993) *J. Chem. Phys.* **98,** 7420–7433.
19. Kolinski, A., Galazka, W. & Skolnick, J. (1996) *Proteins* **26,** 271–287.
20. Plotkin, S. S., Wang, J. & Wolynes, P. G. (1997) *J. Chem. Phys.* **106,** 2932–2948.
21. Doyle, R., Simons, K., Qian, H. & Baker, D. (1997) *Proteins* **29,** 282–291.
22. Sorenson, J. M. & Head-Gordon, T. (1998) *Folding Des.* **3,** 523–534.
23. Chan, H. S. (2000) *Proteins* **40,** 543–571.
24. Vaart, A. van der, Bursulaya, B. D., Brooks, C. L., III, & Merz, K. M., Jr. (2000) *J. Phys. Chem. B* **104,** 9554–9563.
25. Eastwood, M. P. & Wolynes, P. G. (2001) *J. Chem. Phys.* **114,** 4702–4716.
26. Fernández, A., Colubri, A. & Berry, R. S. (2002) *Physica A (Amsterdam, Neth.)* **307,** 235–259.
27. Czaplewski, C., Ripoll, D. R., Liwo, A., Rodziewicz-Motowidlo, S., Wawak, R. J. & Scheraga, H. A. (2002) *Int. J. Quantum Chem.* **88,** 41–55.
28. Fersht, A. R., Matouschek, A. & Serrano, L. (1992) *J. Mol. Biol.* **224,** 771–782.
29. Onuchic, J. N., Wolynes, P. G., Luthey-Schulten, Z. & Socci, N. D. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 3626–3630.
30. Shea, J. E. & Brooks, C. L., III (2001) *Annu. Rev. Phys. Chem.* **52,** 499–535.
31. Daggett, V. & Levitt, M. (1994) *Curr. Opin. Struct. Biol.* **4,** 291–295.
32. Young, W. S. & Brooks III, C. L. (1996) *J. Mol. Biol.* **259,** 560–572.
33. Snow, C. D., Nguyen, H., Pande, V. S. & Gruebele, M. (2002) *Nature* **420,** 102–106.
34. Boczko, E. M. & Brooks, C. L., III (1995) *Science* **269,** 393–396.
35. Duan, Y. & Kollman, P. A. (1998) *Science* **282,** 740–744.
36. Kazmirski, S. L., Wong, K.-B., Freund, S. M., Tan, Y.-J., Fersht, A. R. & Daggett, V. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 4349–4354.
37. Garcia, A. E. & Onuchic, J. N. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 13898–13903.
38. Rhee, Y. M., Sorin, E. J., Jayachandran, G., Lindahl, E. & Pande, V. S. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 6456–6461.
39. Guo, Z. & Thirumalai, D. (1995) *Biopolymers* **36,** 83–102.
40. Zhou, Y. & Karplus, M. (1999) *Nature* **401,** 400–403.
41. Clementi, C., Nymeyer, H. & Onuchic, J. N. (2000) *J. Mol. Biol.* **298,** 937–953.
42. Vendruscolo, M., Paci, E., Dobson, C. M. & Karplus, M. (2001) *Nature* **409,** 641–645.
43. Koga, N. & Takada, S. (2001) *J. Mol. Biol.* **313,** 171–180.
44. Clementi, C., Jennings, P. A. & Onuchic, J. N. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 5871–5876.
45. Onuchic, J. N., Socci, N. D., Luthey-Schulten, Z. & Wolynes, P. G. (1996) *Folding Des.* **1,** 441–450.
46. Du, R., Pande, V. S., Grosberg, A. Yu., Tanaka, T. & Shakhnovich, E. S. (1998) *J. Chem. Phys.* **108,** 334–350.
47. Miyazawa, S. & Jernigan, R. L. (1996) *J. Mol. Biol.* **256,** 623–644.
48. Plaxco, K. W., Simons, K. T. & Baker, D. (1998) *J. Mol. Biol.* **277,** 985–994.
49. Shea, J. E., Onuchic, J. N. & Brooks, C. L., III (1999) *Proc. Natl. Acad. Sci. USA* **96,** 12512–12517.
50. Kaya, H. & Chan, H. S. (2003) *Proteins* **52,** 524–533.
51. Jewett, A. I., Pande, V. S. & Plaxco, K. W. (2003) *J. Mol. Biol.* **326,** 247–253.
52. Horovitz, A. & Fersht, A. (1992) *J. Mol. Biol.* **224,** 733–740.
53. Milet, A., Moszynski, R., Womer, P. E. S. & van der Avoird, A. (1999) *J. Phys. Chem. A* **103,** 6811–6819.
54. Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I. & Baker, D. (1999) *Nat. Struct. Biol.* **11,** 1016–1024.
55. Martinez, J. C. & Serrano, L. (1999) *Nat. Struct. Biol.* **6,** 1010–1016.
56. Chiti, F., Taddei, N., White, P. M., Bucciantini, M., Magherini, F., Stefani, M. & Dobson, C. M. (1999) *Nat. Struct. Biol.* **6,** 1005–1009.
57. Fulton, K. F., Main, E. R. G., Daggett, V. & Jackson, S. E. (1999) *J. Mol. Biol.* **291,** 445–461.
58. Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1995) *J. Mol. Biol.* **254,** 260–288.
59. Kim, D. E., Fisher, C. & Baker, D. (2000) *J. Mol. Biol.* **298,** 971–984.