Available online at www.sciencedirect.com
SCIENCE DIRECT®
Chemical Physics 323 (2006) 66–77
Chemical Physics
www.elsevier.com/locate/chemphys

On the role of chemical detail in simulating protein folding kinetics

Young Min Rhee [a], Vijay S. Pande [a,b,*]

[a] Department of Chemistry, Stanford University, Stanford, CA 94305-5080, United States
[b] Department of Structural Biology, Stanford University, Stanford, CA 94305-5080, United States

Received 11 February 2005; accepted 16 August 2005
Available online 18 October 2005

Abstract

Is an all-atom representation for protein and solvent necessary for simulating protein folding kinetics or can simpler models reproduce the results of more complex models? This question is relevant not just for simulation methodology, but also for the general understanding of the chemical details relevant for protein dynamics. With recent advances in computational methodology, it is now possible to simulate the folding kinetics of small proteins in all-atom detail. Therefore, with both detailed and simplified models of folding in hand, the outstanding questions are what the differences in these models are for the description of protein folding dynamics, and how we can *quantitatively* compare the folding mechanisms found in the models. To address the outstanding problem of how to determine the differences between folding mechanism in a sensitive and quantitative manner, we suggest a new method to quantify the non-linear correlation in folding commitment probability (P_{fold}) values. We use this method to probe the differences between a wide range of models for folding simulations, ranging from coarse grained Gō models to all-atom models with implicit or explicit solvation. While the differences between less-detailed models (Gō and implicit solvation models) and explicit solvation models are large, the differences within various explicit solvation models appear to be small, suggesting that the discrete nature of water may play a role in folding kinetics.

Keywords: Protein folding; Solvent model; Distributed computing; Molecular dynamics

1. Introduction

The study of protein folding dynamics has long been a target for many computational chemists and biologists. Until very recently, folding studies have usually adopted computationally less demanding implicit solvent models [1]. While explicit solvent has also been used, it has typically been restricted to fast unfolding dynamics at very high temperatures [2] or free energy methods [3,4], primarily due to the enormous demands on the processor power. However, with advances in computer technology [5], such as algorithms that can harness the power of new distributed computing technology [6,7], it has now become possible to directly simulate folding dynamics in explicit solvent with tens to hundreds of microseconds of aggregate sampling [8,9]. Accordingly, with the newly found ability to simulate folding with explicit solvent models, we now have the opportunity to investigate how more detailed models affect the nature of the mechanism of folding.

Since the fact that a model is computationally demanding does not guarantee its accuracy [10], a comparison between simulations cannot directly clarify which model is more accurate – such a comparison can only be achieved by a quantitative comparison with experiment. Instead, our goal is to test the role of chemical detail: if simpler models agree with more complex models, it suggests that the details present in complex models do not play a significant role in the folding mechanism and simpler models should be sufficient to understand protein folding. If the models disagree, then we have direct evidence that chemical detail can play a significant role and we would be able to identify the critical elements leading to the differences in these models.

* Corresponding author. Tel.: +1 650 723 3660; fax: +1 650 725 0259.
E-mail address: pande@stanford.edu (V.S. Pande).
URL: http://pande.stanford.edu (V.S. Pande).

0301-0104/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemphys.2005.08.060

There have been a number of reports regarding the comparison between implicit and explicit solvent models. For example, Bursulaya and Brooks applied the free energy landscape approach [4] to a β-sheet protein reporting overall agreements between the two models with minor discrepancies [11]. On the other hand, Zhou and Berne found significant difference for a β-hairpin with the implicit model making erroneous salt bridges between charged residues [12,13]. In this regard, obtaining more detailed and thorough comparisons will be important for understanding the models.

Moreover, these previous works underscore the great challenges involved with a quantitative comparison between solvent models. Unlike other properties relevant to biomolecules, such as the comparison of solvent models for thermodynamics calculations (or solvation free energy) [10,14,15], it is not obvious how to quantitatively compare solvent models in protein folding dynamics. Namely, what quantities must one compare to evaluate whether folding dynamics is similar? In the case of solvation free energy, there is a clear target for quantitative comparison (the solvation free energy itself), and experimental data is usually available as a reference. In folding dynamics, where the materials of comparison are ensembles of trajectories, the means to perform a direct comparison is conceptually unclear. A natural choice is the comparison of folding rates from different models with experiment. The failure to quantitatively predict the rate likely signals some shortcomings in the model, which may imply additional failures beyond the prediction of rates. However, it is also possible that a set of models can agree on the rate of folding but disagree on mechanistic aspects. Indeed, it has been recently shown that the folding rates obtained with an implicit and an explicit solvent models were in agreement even though aspects of the folding mechanism and the characteristics of the transition states were in disagreement [8].

Such an ambiguity might be avoided by using trajectories themselves in the comparison. However, a quantitative comparison of trajectories is not trivial. Protein folding is non-processive: the series of snapshots of a trajectory does not monotonically drift from the unfolded to the folded state. The inherent randomness in this diffusive process is a great obstacle in obtaining a meaningful comparison. Moreover, even if a comparison between two trajectories is obtained, it is not statistically meaningful unless the transition path ensemble is calculated, and obtaining such an ensemble is still a very demanding task [16].

The comparison of folding dynamics from different models could become relatively easy if one can obtain the high dimensional free energy surface of folding in each model. The drawback of this approach lies in the fact that obtaining a reliable free energy surface is another demanding process and one must typically make assumptions of relevant reaction coordinates and calculate the free energy projected onto these coordinates. Indeed, the reaction coordinates chosen could themselves also impact the results [17].

Here, we propose an alternative means to compare folding kinetics. It uses a new method to quantify the non-linear correlation in conformation specific folding probability (P_{fold}). P_{fold} is defined as the probability that a given conformation will commit to the folded state before reaching the unfolded state [18]. As will be shown in the following section, the correlation of P_{fold} from different simulations is usually non-linear, and its observation enables us to predict the change of free energy surface with different models. In this work, we report extensive P_{fold} comparisons between many different protein and solvent models for a small protein folding. By examining a range of different models, such comparisons should reveal the role of detail in the folding mechanism predicted by these folding simulations.

Why would one examine P_{fold} versus other degrees of freedom which may be kinetically relevant? In general, choosing kinetically relevant degrees of freedom is a great challenge [16]. One test of whether a putative reaction coordinate is kinetically relevant is to investigate its correlation with P_{fold} [19,20]. Therefore, while a P_{fold} calculation is very computationally demanding, one can benefit from its examination because P_{fold} is, by construction, a kinetically relevant degree of freedom. Moreover, P_{fold} correlation is a sensitive measure of changes in the mechanism: a shift or broadening of the primary rate limiting free energy barrier for folding will lead to a significant change in P_{fold} values. Thus, if there is a strong correlation in P_{fold}, the kinetics is highly related and the mechanism is essentially identical.

In addition, we stress that P_{fold} correlations may not be necessarily linear. Rhee et al. [8] interpreted a non-linear P_{fold} correlation as a significant deviation in the folding mechanism. Below, we present a simple theory to greatly enhance this interpretation, by deconvoluting issues of simple transformations of the kinetics (e.g., a shift in the transition state) from more complex changes in the dynamics.

2. Methods

2.1. How to quantify the change in folding mechanism: a new scheme for fitting a non-linear P_{fold} correlation

The use of the commitment probability (P_{fold}) to study protein folding dates back to the study by Du et al. [18], where they found that P_{fold} can be chosen as a transition coordinate along which the reaction progresses most slowly. Based on this, it was also shown that the validity of using other geometrical parameters such as the number of native contacts could be verified from their correlations with the transmission probability. Since kinetically relevant degrees of freedom correlate with P_{fold}, it is in fact the natural degree of freedom to examine for comparing the nature of folding kinetics. Indeed, it has been further found that this folding commitment probability has a strong dependence on the shape of the free energy surface [19,21,22], and accordingly, the deviations in P_{fold} with changes in the protein and/or solvent model can be used to map the change in the free energy surface.

In a one-dimensional free energy barrier model of folding, the change of the free energy barrier from using different models is closely related to the non-linearity in the correlation between P_{fold} values (see Fig. 1(a)). In reality, protein folding is a complicated reaction in a multi-dimensional free energy surface, and the correlation will likely have scatter in a band as shown in Fig. 1(b) unless the multi-dimensional free energy surface is uniformly shifted by the use of the different model. Therefore, by observing the correlation between P_{fold} values from different models, one can estimate the effect of using the different models on protein folding dynamics and the related multi-dimensional free energy surface.

In this regard, we wish to develop a means to deconvolute the simple differences in free energy surfaces (such as shifts in the transition state) from more complex differences in kinetics. To do so, we describe the surface with a one-dimensional model of the surface with a simple quadratic free energy barrier $G(x) = -x^2$. In this case, the folding probability in the original surface is given as

$$P_{\text{fold}}(x) = \frac{\int_{-1}^{x} e^{-\beta y^2}\,dy}{\int_{-1}^{1} e^{-\beta y^2}\,dy}. \tag{1}$$

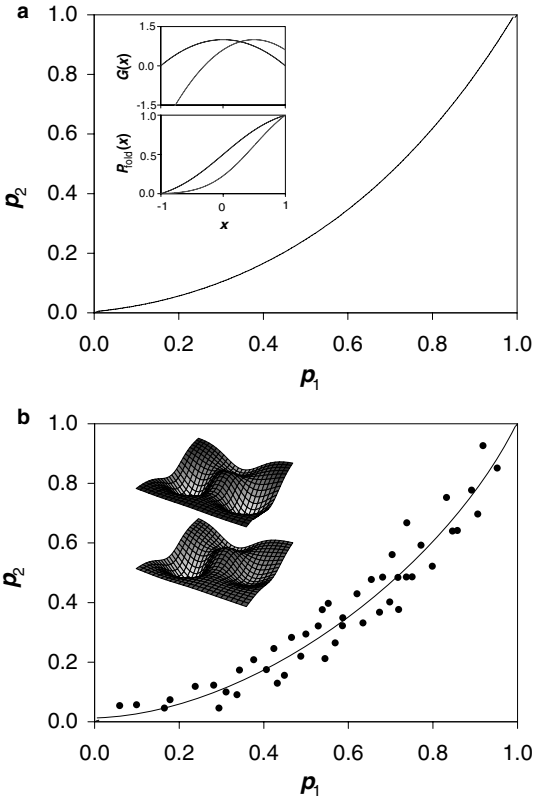

Fig. 1. Schematic illustration P_{fold} correlations between different folding free energy surface: (a) one-dimensional case and (b) multi-dimensional case. Horizontal axis represents P_{fold} from one model (blue surfaces in the insets), and the vertical axis represents P_{fold} from another model (red surfaces in the insets). The one-dimensional case is drawn with actual data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In the first order approximation, the change of potential can be described with the shift of the transition state position x_{TS} and the width of potential σ, or

$$G'(x) = -\left(\frac{x - x_{\text{TS}}}{\sigma}\right)^2. \tag{2}$$

On this new surface, the folding probability is given as

$$P'_{\text{fold}}(x) = \frac{\int_{-1}^{x} e^{-\beta\left(\frac{y-x_{\text{TS}}}{\sigma}\right)^2}\,dy}{\int_{-1}^{1} e^{-\beta\left(\frac{y-x_{\text{TS}}}{\sigma}\right)^2}\,dy}. \tag{3}$$

By fitting the correlation $(P_{\text{fold}}, P'_{\text{fold}})$, one can obtain the change in x_{TS} and σ. Also, the scatter of data is obtained with the root mean square (RMS) deviation of the data from the fit (δ). Physically, these variables are the means to describe the shift of the transition state, change of the barrier width, and the distortion of the free energy barrier, respectively.

2.2. Simulation methods

We compare models by studying the folding of BBA5, a 23-residue mini-protein designed and characterized by Imperiali group [23,24]. We have chosen BBA5 since it is a small, yet stable and structurally well defined protein. Moreover, BBA5 has been recently simulated [8] with the Garcia–Sanbonmatsu modified version (AMBER-GS [25]) of the AMBER94 protein force field [26] and the TIP3P water model [27]. In this work, folding probabilities were calculated for 80 conformations along the reported 13 folding trajectories from the original explicit solvent simulation [8]. The folding probability was also calculated in several other models: TIP4P, M20, and M24 explicit solvation, GB/SA implicit solvation, and coarse grained and hybrid Gō models. The M20 and M24 models are TIP3P variants which have been designed to reduce the error in solvation free energies of amino acid side chain analogs [10] (see Table 1 for parameters). The coarse grained BBA5 Gō model simulations followed the protocol of Clementi and co-workers [28]. Finally, a hybrid Gō/AMBER-GS/TIP3P model was used to test the sensitivity of physical force fields to Gō model interactions; this hybrid model used a Hamiltonian which was a linear sum of the all-atom AMBER-GS model and an all-atom Gō model (see Appendix A for details).

To measure the folding probability of any given conformation, 100 independent molecular dynamics simulations were performed with randomly chosen initial velocities. With 100 samples, the standard deviation in the calculated P_{fold} is 0.05 or less in all cases. For each trajectory, the simulation was continued for 5 ns. After 5 ns, more than 90% of the trajectories committed either to the native or to the folded state. The criteria of folded and unfolded states were the same as in the previous report [8]. Namely, a conformation was considered to be folded if it has both low α-carbon root mean square distance ($\text{RMSD}_{C\alpha}$) from the native structure (<3.1 Å) and well-formed secondary structure (a

Table 1
Parameters and properties of the water models examined

Model	ε (kcal/mol)	σ (Å)	Density (g/cm^3)	ΔH_{vap} (kcal/mol)	Mean error (kcal/mol)	RMS error (kcal/mol)
TIP4P	0.1555	3.1536	0.9997	10.412	0.71	0.82
TIP3P	0.1521	3.1506	0.9859	10.091	0.5	0.64
M20	0.20	3.120	0.9976	10.044	0.18	0.37
M24	0.24	3.111	0.9976	9.948	0.00	0.36
GB/SA*	–	–	–	–	−0.63*	0.98*

* Mean and RMS errors were calculated using Val, Ile, Ser, Asn, Met, and Phe sidechain analogs using data from [33].

hairpin and a helix). A conformation was considered to be unfolded when $RMSD_{C\alpha}$ is high (>4.0 Å) and the secondary structure is broken. These $RMSD_{C\alpha}$ cutoff values were obtained from bimodal $RMSD_{C\alpha}$ distribution with simulations started from the native structure in TIP3P water. Native state boundary is selected as one standard deviation above the average of the low $RMSD_{C\alpha}$ component of the distribution, while the unfolded boundary is chosen as one-standard deviation below the average of the high $RMSD_{C\alpha}$ component (see [8] for details). The total simulation time in this work is ~500 μs, which would require more than 500 CPU years on a modern personal computer. Using a fraction of ~170,000 CPUs on Folding@Home, the entire calculation took approximately one wall-clock month.

The dependence of the folding probability on different models was measured as follows. For implicit solvent models, only the protein molecules were taken from the original simulation boxes. For different explicit solvent models, the protein molecules were taken from the original boxes and re-solvated with a pre-equilibrated box of corresponding solvent molecules. A total of 3938 water molecules were added with a chloride ion in a cubic box of 50 Å length on each side. This re-solvated system was equilibrated using 100 steps of steepest descent energy minimization followed by 100 ps of molecular dynamics. During this equilibration process, the protein coordinates were fixed in space to prevent any conformational change in the protein molecule. This re-solvating approach assumes that P_{fold} is independent of the solvent fluctuations (or equivalently, it assumes that protein degrees of freedom and solvent degrees of freedom are separated). This assumption was successfully tested by calculating the dependence of P_{fold} on solvent fluctuations (see Appendix A).

Simulations with explicit solvent molecules were performed at constant temperature and pressure (298 K, 1 atm) using the GROMACS molecular dynamics suite [5] modified for the Folding@Home [6,7] infrastructure. The temperature and the pressure were controlled by coupling the system to an external heat bath with a relaxation time of 0.5 ps [29]. The electrostatic interactions were treated using the reaction field method [30] with a cutoff of 10 Å, and 10 Å cutoffs with 8 Å tapers were employed for Lennard-Jones interactions. Nonbonded pair lists were updated every 10 steps of molecular dynamics and the integration step size was 2 fs in all simulations. All bonds involving hydrogen atoms were constrained with the LINCS algorithm [31]. Simulations with implicit solvent were conducted using TINKER package [32]. In this simulation set, Langevin dynamics with water-like viscosity ($\gamma = 91$ ps^{-1}) was used together with Still's GB/SA model of solvation [33]. The same simulation protocols were adopted including cutoffs and tapers to eliminate any false discrepancy.

3. Results and discussion

3.1. Universality of kinetic mechanism in explicit solvent models

Fig. 2 presents the P_{fold} correlations obtained from different explicit solvent models. In general, the differences are much smaller in all cases compared to the ones from the implicit simulations, which will be shown below. However, there is clear shift in the transition state location in all models. This shift is more clearly seen in the progressive scanning on the model space with new M20 and M24 models: the transition state is shifted toward the same direction (native state) and the degree of shift is larger for M24. Interestingly, all of these models have stronger water–protein van der Waals interaction through larger Lennard-Jones ε parameters compared to TIP3P model. Therefore, we can infer that the difference in the interfacial behavior of a different solvent model may have an effect on the folding mechanism. We will discuss this in conjunction with the comparison to implicit models later.

To further test the possibility of tighter coordination of water, we have calculated the solvent distribution as a function of the distance from the hydrophobic core surface of the protein. This solvent distribution function (SDF) is equivalent to the radial distribution function of the solvent except that the distance of a solvent shell is determined from the surface of the hydrophobic core atoms. The details of the method to calculate SDF can be found in Appendix A. SDFs obtained with three different starting structures are presented in Fig. 3. Even though the solvent force field parameters are quite different for different models, to our surprise, the shape of SDFs does not vary to a large extent. However, there is a meaningful difference in the first solvation shell located around 1.5 Å from the hydrophobic surface. As was predicted in the above, SDFs from TIP4P and M24 show higher solvent density around the core than that from TIP3P water. Strikingly, the SDF of TIP4P shows a larger deviation than in M24. Namely,

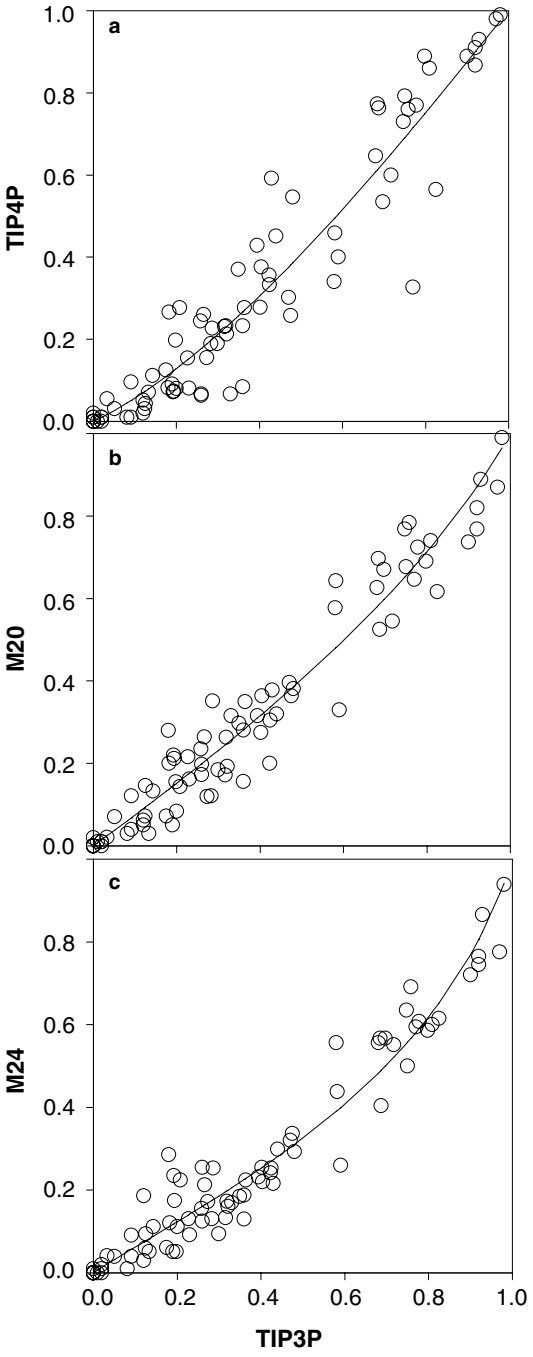

Fig. 2. P_{fold} correlations from simulations using various explicit solvent models. P_{fold} values from (a) TIP4P, (b) M20, and (c) M24 are compared against TIP3P result.

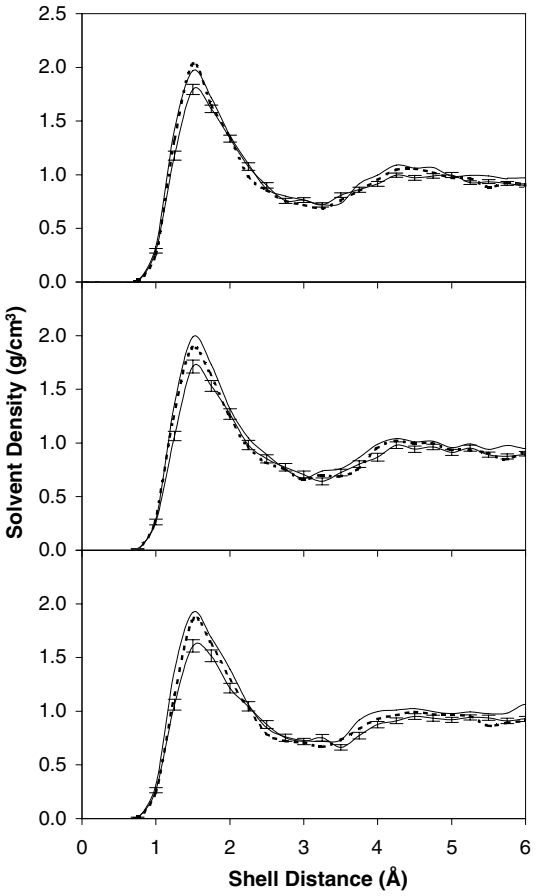

Fig. 3. Solvent distribution functions obtained for three different starting structures used in P_{fold} calculations. Solid black, solid red, and dotted blue lines represent the results from TIP3P, TIP4P, and M24 models of solvent, respectively. Distributions were obtained by averaging 100 independent simulations, and the error bars represent the standard deviation of the average. For visual clarity, error bars are drawn only for the distributions from TIP3P simulations.

the deviation on SDF for TIP4P is similar to or even larger than in the M24 case, while the deviation of the free energy surface measured from the P_{fold} correlation is much more pronounced with M24 water.

What could make P_{fold} values in the TIP4P model so similar to those with TIP3P? One possible answer to this puzzle is the existence of another factor that strongly governs folding dynamics. For example, the changes in viscosity of water from different models can give an explanation to this complication. The viscosities obtained with the periodic perturbation method [34] within our simulation protocols (especially long range interaction cutoffs) are listed in Table 2. It is possible that increased viscosity of TIP4P solvent deters the partially formed hydrophobic core from diffusing out to unfolded state. In such a case, P_{fold} will be generally larger than in M24 as is observed in our results. Indeed, the commitment probability of a diffusive reaction (P_{fold} in this case) is directly related to the diffusion constant of the system [21,35], and the diffusive

Table 2
Shear viscosities for different solvent models*

Model	Viscosity (mPa s)
TIP3P	0.2497
TIP4P	0.4045
M20	0.2312
M24	0.2106

* Computed with 844 water molecules in a 21 Å × 21 Å × 60 Å simulation box. The amplitude of the periodic perturbation was 0.05 nm ps^{-2}. Simulations were performed for 1 ns, and the results during the 0.2–1 ns period were used for averaging purpose.

property of a protein molecule is closely related to the viscosity of the medium.

3.2. Reducing the level of detail: continuum models of water's dielectric and hydrophobic properties (implicit solvent)

Can one reduce the level of detail and still reproduce the detailed nature of the folding mechanism? The natural first step in reducing detail is an implicit solvation model, where only the protein is described in the atomistic details and the solvent is modeled by a continuum field, describing its dielectric and hydrophobic properties. Fig. 4 presents the P_{fold} correlation between explicit (TIP3P) and implicit (GB/SA) solvation models. The difference is quite noticeable: one can see that the transition state in the implicit model is shifted toward the native state considerably, and the distortion of the surface is expected to be significant from the degree of scatters of the data. While the shift in the transition state can be attributed to differences in parameterization of the TIP3P versus GB/SA water models (see Table 1), we attribute the scatter to differences in the description of details between the models. One may argue that this scatter arises from the different solvation free energies with different models. However, as was shown in the previous section, the similarity between TIP3P and M24, which predict significantly different solvation free energies of amino acid side chains, suggests that the scatter will not come from free energetic differences in the models.

Instead, we suggest that the discrete nature of the solvent molecules present in TIP3P (but missing in GB/SA) may also play a role: the geometrical nature of water structure, such as hydrogen bonds [36], is absent in implicit solvation models. This difference in geometry and the accompanied free energetic difference could account for a significant dispersion in the correlation plot. In the transition state of folding, moreover, the protein molecule will have a less compact structure than in the folded state,

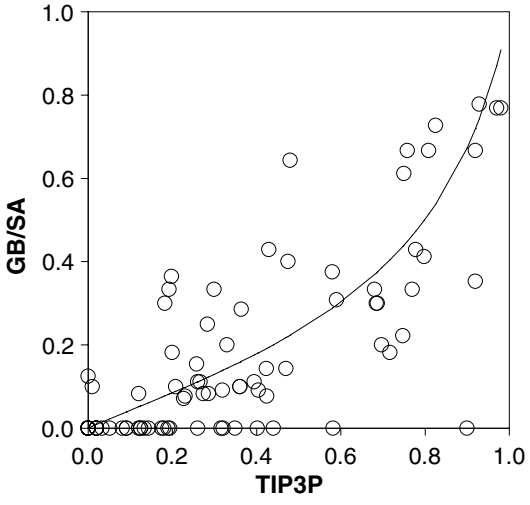

Fig. 4. Comparison of P_{fold} values from simulations using TIP3P and GB/SA solvent models.

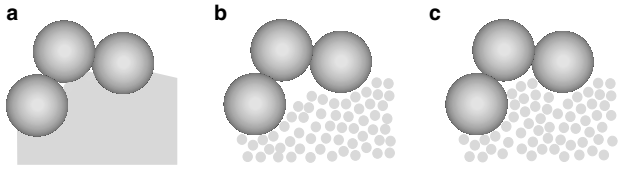

Fig. 5. Schematic illustration of differences in solvent–solute interaction between different solvent models: (a) an implicit solvent model, (b) an explicit solvent model with small solvent–solute attraction, and (c) an explicit model with large solvent–solute attraction.

and a larger hydrophobic surface will be open to contact with the solvent. With an implicit model, the surface is always considered to be solvated, penalizing the relative stability of such a conformation (Fig. 5(a)). On the other hand, explicit solvent molecules can avoid undesirable contacts by simply leaving such an open area as vacuum (Fig. 5(b)) [36,37]. Therefore, a conformation that is considered as a transition state in the explicit model would be less stable in the continuum model, leading to a shift in the transition state location toward the unfolded state as is observed in Fig. 4.

In this regard, it will be intriguing to see the relationship between the protein folding and the solvent distribution around the hydrophobic core residues in the explicit solvent simulations. Namely, with different solvent models, the ability of solvent to generate the void space around the hydrophobic core will change (Figs. 5(b) vs. (c)), and such changes can have an effect on the protein folding dynamics. Indeed, in the previous section, it was inferred that the solvent coordination will be tighter for M24 model compared to TIP3P based on the stronger Lennard-Jones interaction, which was actually confirmed with the SDF shown in Fig. 3. Therefore, the solvent distribution of TIP3P water will be similar to Fig. 5(b) while M24 will be close to the situation in Fig. 5(c). With the stronger interaction between solvent and protein and reduced role of the void space (or discreteness of water) in M24 water, it can be anticipated that the direction of the transition state shift will be the same as in the implicit solvent model. This is indeed the case, as can be clearly seen in Figs. 2 and 4, and we conclude that the solvent coordination around the protein surface affects the protein stability noticeably.

3.3. Reducing the detail even further: Minimalist models of protein folding

Minimalist models, which stress polymeric properties over the specifics of residue–residue interactions, are elegantly simple in their modeling of protein interactions by energetics based on native contacts [38,39]. Moreover, by including a solvent separated minimum in the potential of mean force for protein–protein interactions, a Gō model can recover the discrete nature of water by introducing a desolvation barrier [40] to the protein force field [41]. Indeed, such a model suggested solvent expulsion behavior [41,42], as also seen in explicit solvent simulations [3,43].

To examine this effect, we examined a "hybrid" Gō model, which has Gō-like protein–protein interactions, but surrounded by TIP3P explicit solvent to provide a more physical protein–water interaction (e.g., the discrete nature of water). Since this model has explicit solvent molecules, this hybrid Gō model incorporates not just the solvent separated state of water, but also includes the possibilities for other physical effects associated with explicit water, such as drying and non-pairwise effective protein–protein interactions mediated through water.

Fig. 6 presents the comparison between P_{fold} sets from the AMBER-GS/TIP3P and two different minimalist models. With the coarse grained Gō model, BBA5 did not show any two-state behavior as was reported in other large proteins [28]. In accordance, we used a temperature at which the average of the fraction of native contacts is relatively high (~0.7). Even though a direct comparison is difficult because of the lack of two-state behavior, it can be clearly seen that the correlation is very low (RMS fit error $\delta = 0.26$). When different temperatures were tested, similar correlations were observed. Also, even with the all-atom hybrid model, the correlation is still at a similar level ($\delta = 0.22$).

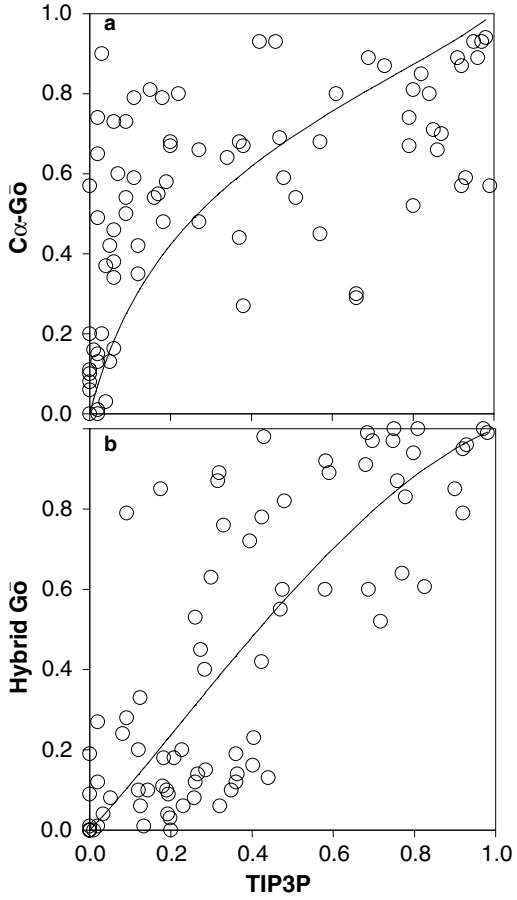

Fig. 6. Comparison of P_{fold} values from simulations using different levels of details in the simulation: (a) a coarse grained Gō model and (b) a hybrid Gō model against all-atom simulation with TIP3P solvent.

Therefore, we infer that the removal of detail in minimalist models (such as the lack of non-native interactions) can significantly alter the folding kinetics from that found in all-atom, explicit solvation models. We stress that the disagreement between these models alone cannot imply that one is "incorrect" – such a judgment can only be reached with quantitative comparisons to experiments. However, the inability of the Gō models employed here to recapitulate the results of more detailed models suggests that atomistic detail does play a significant role, at least for simulations of small proteins such as BBA5. It is possible that Gō models may be more similar to detailed models in larger (e.g., 50–100 residue) proteins, where polymeric and topological properties are likely more significant. Unfortunately, due to the difficulty in simulating such proteins in all-atom detail, it is currently impossible to test that case.

3.4. Further discussion on the comparison of models

How to compare the kinetics of different models is an intrinsically difficult task. The use of P_{fold} can address issues of the quality of the kinetic relevance of the reaction coordinate employed in the analysis, but there other complications remain. In particular, even with an ideal reaction coordinate, how would one use the placement along the reaction coordinate to compare kinetics? Here, we discuss two metrics which yield different information regarding the nature of the differences in the kinetics of the models.

First, a natural metric to quantify the nature of the correlation in P_{fold} values is to examine the scatter of the fit (Fig. 7). A small scatter (such as in the comparison between TIP3P and M24 for example) reflects the intrinsic similarity in the ordering of the conformations along the folding reac-

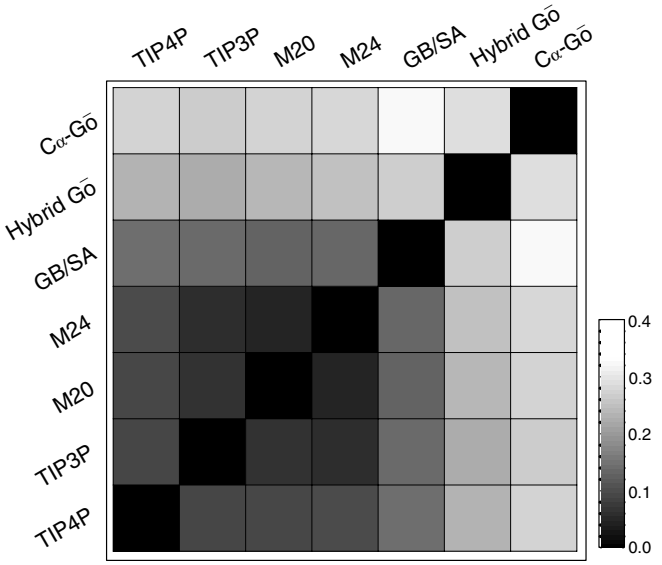

Fig. 7. Similarities between models with different levels of details in the simulation. The similarity is measured with the RMS fit errors in the P_{fold} correlations.

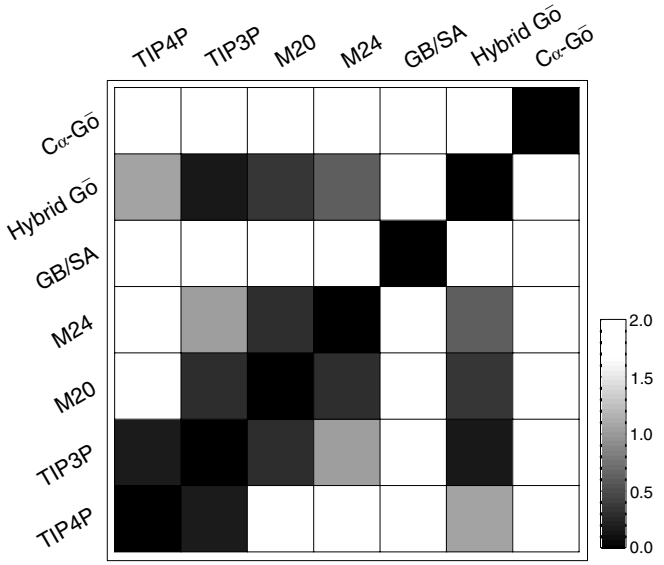

Fig. 8. Similarities between models with different levels of details in the simulations. The similarity is measured with the shift in the transition state location from the fit in the P_{fold} correlation.

tion, even in spite of a major "tilting" of the free energy landscape.

The nature of the shift of the transition state (Fig. 8) is a different metric for comparison between models than the scatter of the fit. It is possible for two models to agree in the ordering of specific states along the reaction pathway (low scatter) but differ by some overall shift in the TS. We see this in comparing TIP3P to the M2x models. The M2x models were built by altering the nature of protein–solvent attraction in order to reparameterize TIP3P to better reproduce solvation free energy properties of small molecules. Thus, one would expect that this would change the stabilization of the folded state as well as shift the transition state. This is seen in the gradual difference in TS location in Fig. 8.

It is also possible for two models to have a small shift in TS, but large scatter, due to differences in the details of folding, but not some very general overall properties (i.e., the location of the TS along the reaction coordinate). This scenario is demonstrated by comparing TIP3P to the hybrid Gō model. The hybrid Gō model uses a combination of an AMBER physical force field, a Gō model, and TIP3P explicit solvent. The transition state location of the hybrid Gō model is closest to TIP3P and has a similar TS shift relationship to the M2x models as AMBER in TIP3P.

4. Conclusions

Is an all-atom, explicit solvent representation needed to faithfully reproduce protein folding kinetics? The answer to that question lies in a tight, quantitative connection to experiment, and unfortunately, current experiments may not provide sufficient detail to resolve this issue. In this paper, we have addressed a related question: what is the role of chemical detail in protein folding simulations and does detail change the nature of the folding mechanism?

Our results suggest that the chemical detail present in all-atom models do play a significant role in the kinetic mechanism, at least for small proteins like BBA5. Comprehensive comparisons between different models are presented in Fig. 9. Also, more quantitative comparisons of the scatter and the transition state shift can be found in Figs. 7 and 8 respectively, where similarities of the folding dynamics between different pairs of models are tabulated in a diagrammatic fashion. We find that the P_{fold} correlation between all-atom models with explicit solvation displayed similar kinetics, perhaps with an overall shift in the transition state. Use of implicit solvation diminished the similarity, and Gō models showed a much more significant deviation.

We stress that the P_{fold} correlation examination is a particularly sensitive test of the similarity of folding kinetics, since P_{fold} varies significantly with even small changes in the underlying free energy landscape. Thus, in interpreting our results, one should keep in mind that for rates and rough aspects of the folding mechanism, explicit (TIP3P) and implicit (GB/SA) water models yielded similar results, even though there was a noticeable deviation in the P_{fold} correlation. Thus, it is certainly possible that current experimental methods for deducing the nature of the folding mechanism may not be able to arbitrate the differences between TIP3P and GB/SA solvent models, for example, and further development of experimental methods would be needed. Indeed, from the agreement in rates and overall mechanism in TIP3P and GB/SA simulations of BBA5 folding, we should still consider the possibility that these differences may be relatively unimportant. Various degrees of agreement and disagreement between different models in earlier studies may be explained in the same manner: differences can be found only when a sensitive measure is used.

However, the differences between the Gō models and the explicit and more traditional implicit solvation models were significantly larger. This is in agreement with the findings by Cavalli et al., where the free energy surfaces and the related properties from implicit solvent model and all-atom Gō model were found to be significantly different [44]. This reflects a much more significant difference in the folding mechanisms of BBA5 in these models. We stress that this fact alone only speaks to the differences between these models and cannot comment on their experimental validity. Toward this end, it is natural to examine the ability of different models to predict multiple experimental properties, such as rates [8,28], free energies [14], and other quantitative observables. Finally, it is also possible that our results are only valid for small proteins and that the details may be less involved in larger proteins (e.g., 50–100 residues). In this regard, it will be also interesting to examine

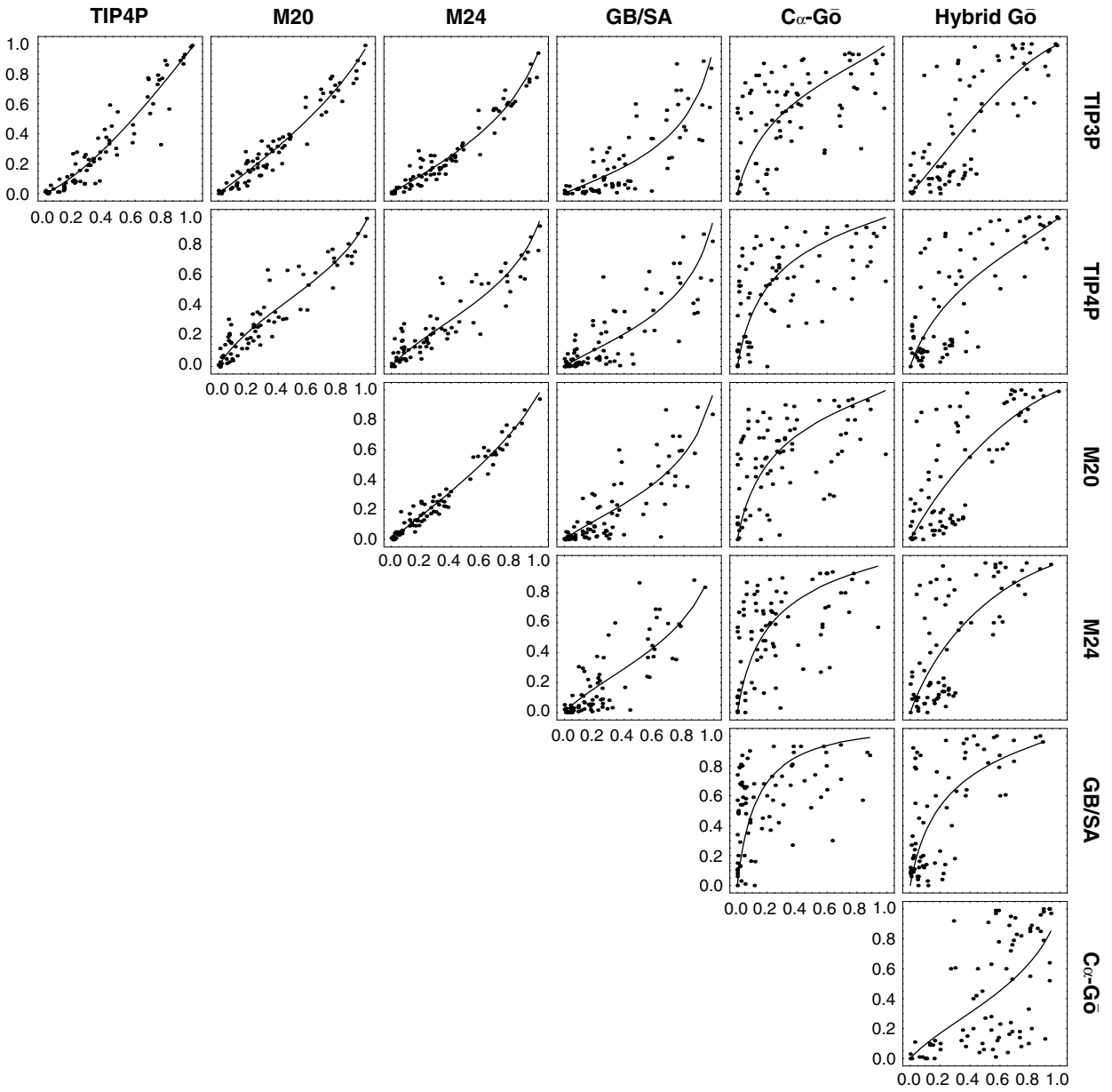

Fig. 9. Comparisons of P_{fold} values from various models of protein and solvent. Filled circles are actual data from simulation. Solid lines represent the fits using one-dimensional free energy models.

large systems as a continuation of this study. Clearly, such a study is a natural direction for future work.

In conclusion, whether P_{fold} is too sensitive or not sensitive enough will largely rest on the quantities of interest to be predicted from simulation. If one only cares about rates and other macroscopic properties, a strong P_{fold} correlation is likely not important. Also, if one is largely interested in rough mechanistic properties, a rough correlation may be sufficient (for example, the mechanism of BBA5 folding in TIP3P and GB/SA were similar [8,45] and a reasonably strong P_{fold} correlation was found). However, if one cares about specific details of the mechanism, P_{fold} may not be sensitive enough (for example, there were only minor differences in the P_{fold} correlation between different explicit solvation models). Indeed, if the folding mechanism of two models is similar to atomic detail, the ordering of the states along the reaction pathway (which is what P_{fold} reports) would be constrained to be similar, and hence would result in a P_{fold} strong correlation. In this case, likely more strict criteria will be needed to not merely compare the ordering of states along the reaction pathway, but more specific structural elements of the mechanism itself. Nevertheless, we have demonstrated here that P_{fold} is sufficiently sensitive to distinguish between simple coarse grained models (Gō models), a physically based implicit solvent model (GB/SA), and more detailed explicit solvent models (TIP3P, etc), directly demonstrating its sensitivity amongst the models currently used to simulate protein folding.

Acknowledgements

This work was supported by NSF Molecular Biophysics (MCB-0317072). Y.M.R. acknowledges a support from William Nichols Fellowship.

Appendix A

A.1. Separation of the solvent degrees of freedom from the protein degrees of freedom

To measure the folding probability at a given protein conformation with different solvent models, it is necessary to show that the solvent degrees of freedom do not change the folding probability. Namely, because it is necessary to re-equilibrate the system with a different solvent model, it is required to show that such a re-equilibration does not alter the folding probability of the protein. In fact, the solvent degrees of freedom were found to be important in conformational changes of small molecules such as ion pair dissociation [46] and alanine dipeptide isomerization [47], where the changes of solvent conformations resulted in dramatic changes in reaction probabilities.

To test this possibility in our system, we computed the P_{fold} distribution with many different solvent conformations. For a selected conformation (protein and solvent) from folding simulations using TIP3P water [8], a new molecular dynamics simulation was initiated and continued for 10 ns with the protein conformation frozen in space. Conformations were sampled at an interval of every 100 ps, and the resulting 100 conformations were used to obtain P_{fold}. If the solvent degrees of freedom are not a decisive factor in the folding dynamics, all the trials will have the same native probability of folding, and the following distribution of P_{fold} will be binomial (or Gaussian if the sampling is large enough). This approach was performed for a number of different protein conformations with a range of averages P_{fold} values.

Indeed, the calculated P_{fold} distributions exactly match the binomial distributions as shown in Fig. A.1. The same degree of agreement was found for all conformations tried in a wide range of P_{fold} values. Therefore, we conclude that the solvent degrees of freedom are not a decisive factor for the "fate" of this protein in the course of folding. Accordingly, we assume our scheme of stripping the protein conformation from TIP3P trajectory and then re-solvating it with new solvent conformations for different water models is reasonable.

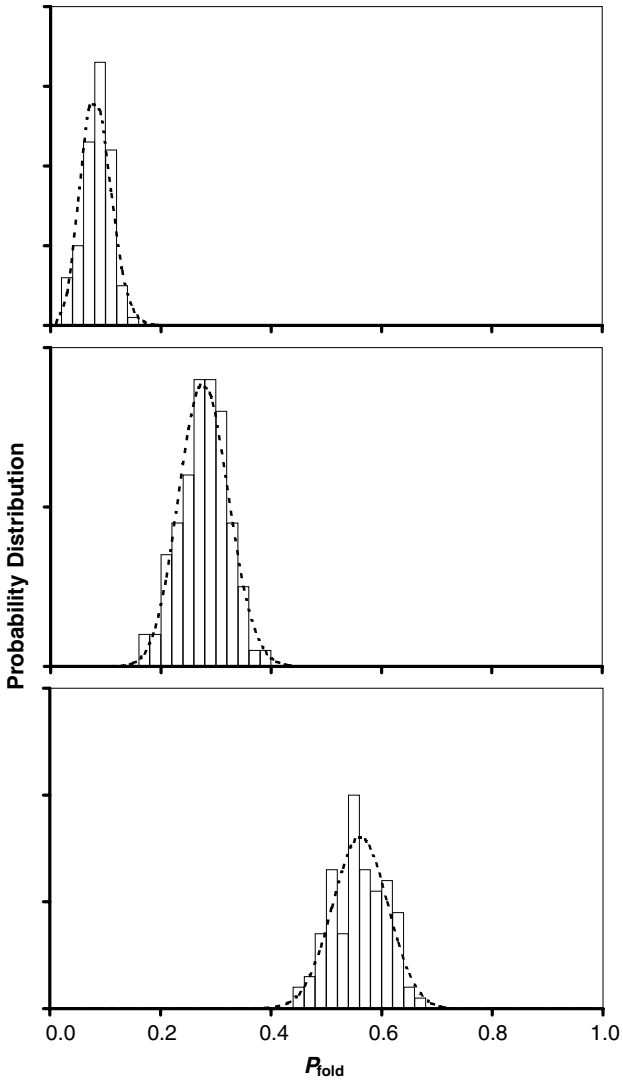

Fig. A.1. Dependence of folding probability on the solvent fluctuations at various protein conformations ranging from almost unfolded (top) to largely folded (bottom). Bars represent the histograms measured from simulations with 100 different solvent conformations. Dashed lines represent the binomial distribution with trial number 100 at the predicted probability. All simulations were performed with TIP3P water.

A.2. Calculation of solvent distribution function

To obtain the solvent density around the hydrophobic core, the following method is employed. A space within a distance r from the hydrophobic core is defined as

$$S(r) = \bigcup_{i}^{N_h} S(r + r_{\text{vdW}}^{i}, \boldsymbol{x}_i) - \bigcup_{j}^{N_p} S(r_{\text{vdW}}^{j}, \boldsymbol{x}_j), \tag{A.1}$$

where $S(a, \boldsymbol{x})$ is a sphere of radius a located at $\mathbf{x}$, and $r_{\text{vdW}}{}^{i}$ and $\mathbf{x}_i$ are the van der Waals radius and the position of the i-th atom in the protein, respectively. Here, the first sum runs over the hydrophobic groups while the second sum runs over all protein atoms. If the volume of $S(r)$ is denoted as $V(r)$ and the number of water molecules in it is $n(r)$, SDF can be determined as

$$\text{SDF}(r) = \frac{\mathrm{d}n}{\mathrm{d}V} = \frac{n(r + \delta r) - n(r)}{V(r + \delta r) - V(r)}. \tag{A.2}$$

Practically, this can be achieved with a grid-based method. Namely, the simulation box was divided into three dimensional grids, and $V(r)$ was obtained by counting the number of grids satisfying Eq. (A.1). In this work, grid size of 0.25 Å was used. Also, the position of the oxygen atom

was used as the position of a given water molecule. From the nature of solvent and from the numerical differentiation applied, SDF obtained in this way showed a relatively large fluctuation. For a given protein conformation, a smooth SDF was obtained by taking an average from 100 independent water conformation sets, which were recorded 200 ps after the P_{fold} simulations were initiated. The error in SDF was estimated as the standard deviation of this average.

A.3. Construction of the hybrid Gō model

An all-atom Gō model is built as follows. In a Gō model, the energy of a protein is directly related to its degree of "nativeness" of a given structure. Usually, the native state is stabilized through additional non-bonded interactions for native contacts and modifications of the backbone dihedral potentials to favor native conformation [48]. Accordingly, we have modified the Lennard-Jones (LJ) interaction as

$$E_{\text{LJ-Go}}(r_{ij}) = \begin{cases} \frac{C_{12}^{ij}}{r_{ij}^{12}} - \frac{C_6^{ij}}{r_{ij}^6} & i-j: \text{ close contact}, \\ \varepsilon_a \left(\frac{C_{12}^{ij}}{r_{ij}^{12}} - \frac{C_6^{ij}}{r_{ij}^6} \right) & i-j: \text{ native contact}, \\ \varepsilon_b \frac{C_{12}^{ij}}{r_{ij}^{12}} & i-j: \text{ non-native contact}, \end{cases} \tag{A.3}$$

where C_6^{ij} and C_{12}^{ij} are the original LJ potential parameters taken from the AMBER-GS force field [25]. Any atom pair is considered to be in a close contact when they are separated by less than three amino acid residues. For other non-close atom pairs, a native contact is declared for atoms within 5.5 Å distance in the native structure. The remaining atom pairs are considered to be in non-native contacts. With this modification, attractions between pairs of native contacts are amplified by a factor ε_a ($\varepsilon_a > 1$), while attractions in non-native interactions are ignored. The factor ε_b ($\varepsilon_b < 1$) prevents the potential from being excessively repulsive at short distances. Dihedral potentials of the backbone φ and ψ angles are also replaced by

$$E_{\text{dih-Go}}(\phi_i) = k_\phi[1 - \cos(\phi_i - \phi_{i,0})], \tag{A.4}$$

where $\phi_{i,0}$ is the ith φ or ψ angle in the native conformation. Because the purpose of the perturbation is to force the native state to be stabilized, it is unnecessary to change any interactions involving solvent molecules. Thus, our Gō model has three adjustable parameters ε_a, ε_b, and k_ϕ, which were chosen as 4.0, 0.1, and 80 kJ/mol, respectively.

This all-atom Gō model H_{Go} is combined with the unperturbed physical potential H_0 to create the hybrid Hamiltonian H:

$$H = (1-\lambda)H_0 + \lambda H_{\text{Go}}. \tag{A.5}$$

Here, λ can gradually dial the system from a full physical model H_0 at $\lambda = 0$ to pure Gō model at $\lambda = 1$. We chose to use $\lambda = 0.05$, at which the protein native structure was considerably more stable than in H_0 with a 100-fold increase in the folding rate.

References

[1] C.D. Snow, E.J. Sorin, Y.M. Rhee, V.S. Pande, Annu. Rev. Biophys. Biomol. Struct. 34 (2005) 43.
[2] A.R. Fersht, V. Daggett, Cell 108 (2002) 573.
[3] F.B. Sheinerman, C.L. Brooks, J. Mol. Biol. 278 (1998) 439.
[4] J.E. Shea, C.L. Brooks, Annu. Rev. Phys. Chem. 52 (2001) 499.
[5] E. Lindahl, B. Hess, D. van der Spoel, J. Mol. Model. 7 (2001) 306.
[6] M. Shirts, V.S. Pande, Science 290 (2000) 1903.
[7] V.S. Pande, I. Baker, J. Chapman, S.P. Elmer, S. Khaliq, S.M. Larson, Y.M. Rhee, M.R. Shirts, C.D. Snow, E.J. Sorin, B. Zagrovic, Biopolymers 68 (2003) 91.
[8] Y.M. Rhee, E.J. Sorin, G. Jayachandran, E. Lindahl, V.S. Pande, Proc. Natl. Acad. Sci. USA 101 (2004) 6456.
[9] G. Jayachandran, V. Vishal, V.S. Pande, in submission.
[10] M.R. Shirts, V.S. Pande, J. Chem. Phys. 122 (2005) 134508.
[11] B.D. Bursulaya, C.L. Brooks, J. Phys. Chem. B 104 (2000) 12378.
[12] R.H. Zhou, B.J. Berne, Proc. Natl. Acad. Sci. USA 99 (2002) 12777.
[13] R.H. Zhou, Proteins 53 (2003) 148.
[14] M.R. Shirts, J.W. Pitera, W.C. Swope, V.S. Pande, J. Chem. Phys. 119 (2003) 5740.
[15] M. Feig, A. Onufriev, M.S. Lee, W. Im, D.A. Case, C.L. Brooks, J. Comput. Chem. 25 (2004) 265.
[16] P.G. Bolhuis, D. Chandler, C. Dellago, P.L. Geissler, Annu. Rev. Phys. Chem. 53 (2002) 291.
[17] V.S. Pande, A.Y. Grosberg, T. Tanaka, D.S. Rokhsar, Curr. Opin. Struct. Biol. 8 (1998) 68.
[18] R. Du, V.S. Pande, A.Y. Grosberg, T. Tanaka, E.S. Shakhnovich, J. Chem. Phys. 108 (1998) 334.
[19] Y.M. Rhee, V.S. Pande, J. Phys. Chem. B 109 (2005) 6780.
[20] A. Ma, A.R. Dinner, J. Phys. Chem. B 109 (2005) 6769.
[21] C.W. Gardiner, Handbook of Stochastic Methods, Springer, Berlin, 1985.
[22] G. Hummer, J. Chem. Phys. 120 (2004) 516.
[23] M.D. Struthers, R.P. Cheng, B. Imperiali, Science 271 (1996) 342.
[24] M. Struthers, J.J. Ottesen, B. Imperiali, Fold. Des. 3 (1998) 95.
[25] A.E. Garcia, K.Y. Sanbonmatsu, Proc. Natl. Acad. Sci. USA 99 (2002) 2782.
[26] W.D. Cornell, P. Cieplak, C.I. Barly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J. Caldwell, P.A. Kollman, J. Am. Chem. Soc. 117 (1995) 5179.
[27] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, M.L. Klein, J. Chem. Phys. 79 (1983) 926.
[28] L.L. Chavez, J.N. Onuchic, C. Clementi, J. Am. Chem. Soc. 126 (2004) 8426.
[29] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, A. Dinola, J.R. Haak, J. Chem. Phys. 81 (1984) 3684.
[30] M. Neumann, O. Steinhauser, Mol. Phys. 39 (1980) 437.
[31] B. Hess, H. Bekker, H.J.C. Berendsen, J.G.E.M. Fraaije, J. Comput. Chem. 18 (1997) 1463.
[32] J.W. Ponder, TINKER, Software Tools for Molecular Design, Department of Biochemistry and Molecular Biophysics, Washington University, St. Louis, 2000.
[33] D. Qiu, P.S. Shenkin, F.P. Hollinger, W.C. Still, J. Phys. Chem. A 101 (1997) 3005.
[34] B. Hess, J. Chem. Phys. 116 (2002) 209.
[35] M. Tachiya, J. Chem. Phys. 69 (1978) 2375.
[36] D. Chandler, Nature 417 (2002) 491.
[37] P.R. ten Wolde, D. Chandler, Proc. Natl. Acad. Sci. USA 99 (2002) 6539.
[38] H. Taketomi, Y. Ueda, N. Go, Int. J. Peptide Protein Res. 7 (1975) 445.
[39] T. Head-Gordon, S. Brown, Curr. Opin. Struct. Biol. 13 (2003) 160.
[40] G. Hummer, S. Garde, A.E. Garcia, M.E. Paulaitis, L.R. Pratt, Proc. Natl. Acad. Sci. USA 95 (1998) 1552.

[41] M.S. Cheung, A.E. Garcia, J.N. Onuchic, Proc. Natl. Acad. Sci. USA 99 (2002) 685.
[42] J. Karanicolas, C.L. Brooks, Protein Sci. 11 (2002) 2351.
[43] J.E. Shea, J.N. Onuchic, C.L. Brooks, Proc. Natl. Acad. Sci. USA 99 (2002) 16064.
[44] A. Cavalli, M. Vendruscolo, E. Paci, Biophys. J. 88 (2005) 3158.
[45] C.D. Snow, N. Nguyen, V.S. Pande, M. Gruebele, Nature 420 (2002) 102.
[46] P.L. Geissler, C. Dellago, D. Chandler, J. Phys. Chem. B 103 (1999) 3706.
[47] P.G. Bolhuis, C. Dellago, D. Chandler, Proc. Natl. Acad. Sci. USA 97 (2000) 5877.
[48] C. Clementi, H. Nymeyer, J.N. Onuchic, J. Mol. Biol. 298 (2000) 937.