PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

An all-atom structure-based potential for proteins: Bridging minimal models with all-atom empirical forcefields

Paul C. Whitford,[1] Jeffrey K. Noel,[1] Shachi Gosavi,[1] Alexander Schug,[1] Kevin Y. Sanbonmatsu,[2] and José N. Onuchic[1]*

[1] Center for Theoretical Biological Physics and Department of Physics, University of California at San Diego, La Jolla, California 92093

[2] Theoretical Biology and Biophysics, Theoretical Division, Los Alamos National Laboratory, MS K710, Los Alamos, New Mexico 87545

ABSTRACT

Protein dynamics take place on many time and length scales. Coarse-grained structure-based (Gō) models utilize the funneled energy landscape theory of protein folding to provide an understanding of both long time and long length scale dynamics. All-atom empirical forcefields with explicit solvent can elucidate our understanding of short time dynamics with high energetic and structural resolution. Thus, structure-based models with atomic details included can be used to bridge our understanding between these two approaches. We report on the robustness of folding mechanisms in one such all-atom model. Results for the B domain of Protein A, the SH3 domain of C-Src Kinase, and Chymotrypsin Inhibitor 2 are reported. The interplay between side chain packing and backbone folding is explored. We also compare this model to a C_α structure-based model and an all-atom empirical forcefield. Key findings include: (1) backbone collapse is accompanied by partial side chain packing in a cooperative transition and residual side chain packing occurs gradually with decreasing temperature, (2) folding mechanisms are robust to variations of the energetic parameters, (3) protein folding free-energy barriers can be manipulated through parametric modifications, (4) the global folding mechanisms in a C_α model and the all-atom model agree, although differences can be attributed to energetic heterogeneity in the all-atom model, and (5) proline residues have significant effects on folding mechanisms, independent of isomerization effects. Because this structure-based model has atomic resolution, this work lays the foundation for future studies to probe the contributions of specific energetic factors on protein folding and function.

Proteins 2009; 75:430–441.

Key words: energy landscape theory; protein folding; structure-based model; all-atom model; side chain packing.

INTRODUCTION

In recent years the energy landscape theory of protein folding[1–5] has been validated through its application to protein folding,[6–10] oligomerization,[11–14] functional transitions,[15–20] and structure prediction.[21,22] The theory states that proteins are minimally frustrated, that their energy landscape is funnel shaped and that the folded state of the protein is at the bottom of the funnel. Because of the shape of the landscape there is a strong energetic bias towards the folded state of the protein with relatively infrequent trapping caused by non-native interactions. The resulting heterogeneity observed during folding is due to the geometric constraints of the native structure. Thus, models of proteins that have only the native structure encoded have had great success in determining folding mechanisms. Until recently, most models tended to be coarse-grained, which are very useful in understanding global folding dynamics. In commonly used structure-based (Gō) potentials,[9] each residue is represented by a bead centered at the location of the C_α atom [Fig. 1(b)] and only native interactions are stabilizing.

On the other end of the spectrum of structural and energetic details are the computationally intensive all-atom empirical forcefields.[25–30] These forcefields include an atomistic representation of a protein either with an implicit or an explicit solvent. In these potentials, the parameters which determine the interaction between atoms, such as partial

This work was supported by the Center for Theoretical Biological Physics sponsored by the NSF (Grant PHY-0822283) with additional support from NSF-MCB-0543906.
Grant sponsor: National Institutes of Health Molecular Biophysics Training Program at University of California at San Diego; Grant number: T32GM08326; Grant sponsor: NIH; Grant number: R01-GM072686; Grant sponsor: LANL LDRD Program.
*Correspondence to: José N. Onuchic, Center for Theoretical Biological Physics and Department of Physics, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093. E-mail: jonuchic@ctbp.ucsd.edu
Received 11 June 2008; Accepted 8 August 2008
Published online 2 September 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22253

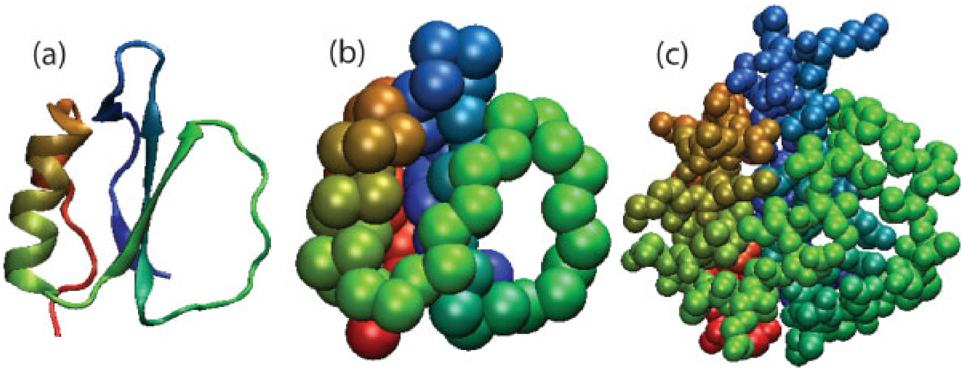

Figure 1

CI2 (Protein Data Base Entry 1YPA[23] shown in (**a**) cartoon representation, (**b**) C_α representation, and (**c**) all-atom (AA) representation. Structures are colored red (C-terminus) to blue (N-terminus). The size of the atoms in the C_α and AA representations correspond to the excluded volume radii used in the C_α[9] and AA models studied in this article. Structures visualized using VMD.[24]

charges and van der Waals radii, are fit to experimental measurements and quantum mechanical calculations. With accurate calibration, a single parameter set may be applied to any protein and, with sufficient computing resources, the dynamics of a protein can be calculated on a computer. The physics-based representation of atom–atom interactions automatically includes electrostatic interactions as well as any non-native interactions that may be present. In principle, these models render knowledge of a native structure unnecessary. A major limitation of these potentials is that they are often too expensive to fold all but small proteins.[30–39] The timescales that can currently be calculated vary from hundreds of nanoseconds to microseconds, depending on the size of the protein. Biological timescales are usually several orders of magnitude larger and these dynamics cannot be accessed using all-atom empirical forcefields. In addition, sensitivity analysis of the dynamics to the parameters is not possible with these all-atom empirical forcefields.

In all-atom empirical forcefields an observed specificity of (i.e., preference for) native interactions is seen as a consequence of many energetic contributions. Because of the complex formulation of these potentials, it is impossible to partition geometric effects from energetic ones. There is a similar restriction in coarse-grained models because of their simplicity. Partitioning these effects is often impossible because geometry is included implicitly through energetic interactions. By studying all-atom models with structure-based potentials,[40–44] because atomic geometry is explicitly included, we can ask to what extent energetics contribute to the apparent native specificity in protein structure, folding, and function. In contrast to enzyme catalysis where specific atomic interactions directly control the chemical reactions, in most cases the energetic specificity required in protein folding is less stringent.

Providing a complete picture of specificity in protein folding and function will require the study of many proteins and many parametric variations. In this article, we lay the foundation for this line of investigation through systematic characterization of a completely specific (only, and all, native interactions are stabilizing) AA structure-based model. We study the effect of varying the parameters of the model on folding barriers, mechanisms, contact formation, and side chain dynamics. The test proteins, B domain of Protein A, SH3 domain of C-Src Kinase, and Chymotrypsin Inhibitor 2 (CI2) (Fig. 2) have been experimentally[47–49] and computationally[8,50–52] well characterized. Additionally, they possess two-state folding dynamics and represent different secondary and tertiary structures. The present model is energetically unfrustrated, with an explicit representation of all non-hydrogen atoms and homogeneous interaction strengths. We find that the folding mechanisms in the model are robust to parameter changes and dynamics agrees well with both the C_α model and an all-atom empirical forcefield with explicit solvent. Further, side chain ordering can be probed explicitly and the effect of prolines can be calculated. This study and model will serve as a basis for future AA models which incorporate nonspecific contributions of energetic frustration, electrostatics, and hydration.

RESULTS

Folding mechanisms are robust to parameter changes

We use a model where the potential energy function is defined by the native state and all heavy (non-hydrogen) atoms are explicitly represented. Any two atoms that are close in the native structure are said to form a native contact. We describe the folding process by using the

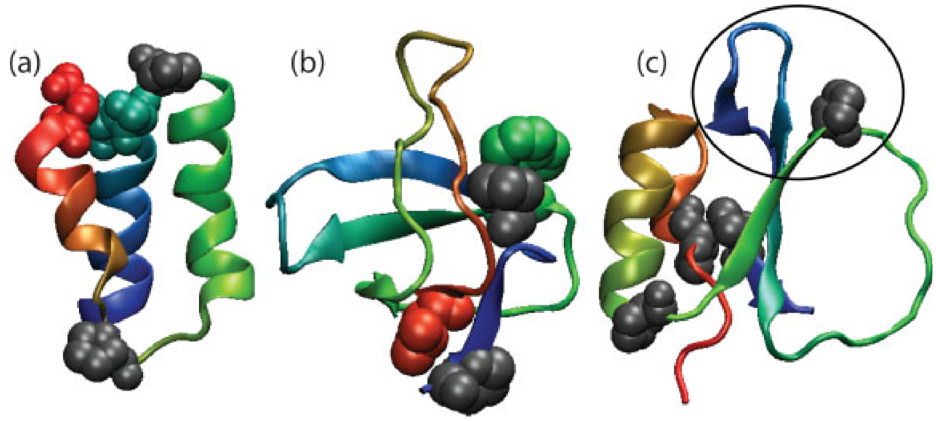

Figure 2

Structures of (**a**) Protein A, (**b**) SH3, and (**c**) CI2 (PDB entries 1BDD,[45] 1FMK,[46] and 1YPA[23] colored red (C-terminus) to blue (N-terminus). These three proteins represent differing structural content and topological complexity. Protein A is a three-helix bundle, SH3 is composed of multiple β strands, and in CI2 an alpha helix flanks a β sheet. Proline residues are shown as gray spheres. In Protein A, Gln1 and Ser31 are shown as colored spheres. In SH3, Val4 and Trp35 are shown as spheres. The mini-core of CI2 is circled.

fraction of native residue pairs in contact Q_{AA} (see "Methods" section). Figure 3(a) shows Q_{AA}, Q_{CA} (fraction of C_α contacts, see "Methods" section) and radius of gyration R_g as functions of time for an AA simulation of CI2, near folding temperature. Because Q_{AA} captures the same collapse events as R_g and Q_{CA} [Fig. 3(b)], Q_{AA} is an useful measure of backbone folding in addition to side chain packing.

It is crucial to understand the parameter dependence of a model before it can be used to make reliable predictions of folding mechanisms. The robustness of the folding mechanism is probed here by characterizing Protein A, SH3, and CI2 for variants of the AA structure-based energy function. Because of the debate about the balance between secondary and tertiary interactions, we vary the ratio of nonlocal contact energy to dihedral angles $R_{C/D}$ and the relative strength of backbone dihedral angles to side chain dihedral angles $R_{BB/SC}$ (see "Methods" section).

To characterize the folding mechanism for different parameter sets we computed the probability of contacts formed as a function of the folding process $P(Q_i,Q_{AA})$. $P(Q_i,Q_{AA})$ is the probability that the contacts involving residue i, Q_i, are formed as a function of Q_{AA}. $P(Q_i,Q_{AA})$ was calculated for the three proteins for 16 different parameter sets (all combinations of $R_{C/D}$ = 1.0,2.0, 3.0,4.0 and $R_{BB/SC}$ = 1.0, 2.0, 3.0,4.0). Figure 4 shows the folding mechanisms for four parameter sets. The difference in folding mechanism between parameter sets i and j can be quantified by the root mean squared deviation in $P(Q_i,Q_{AA})$ over all Q_{AA} and Q_i, ($P_{rms} = \sqrt{\langle (P_i(Q_i,Q_{AA}) - P_j(Q_i,Q_{AA}))^2 \rangle}$). The largest values of P_{rms} for Protein A, SH3, and CI2 were 0.057, 0.097, and 0.077. SH3 is a complicated fold, Protein A a simple fold, and CI2 an intermediate fold.[53] Thus, it is not surprising that energetic modifications have the largest effects on Protein A and the smallest effects on SH3.

Figure 4 shows proline containing regions are less stable to parametric modifications. Regions with prolines, and regions interacting with prolines, form structure earlier (at lower Q) with increased contact strength. This is because contact strength is increased at the expense of dihedral strength. Prolines possess a covalent C_δ—N bond, which limits the mobility of the ϕ dihedral. Removing energy from the dihedrals does not increase flexibility in prolines. However, adding energy to contacts increases structure formation around prolines. For this reason, increasing $R_{C/D}$ stabilizes and promotes earlier formation of proline containing regions.

Fully folded backbone allows for disordered side chains

Although Q_{AA} and Q_{CA} capture the same cooperative folding events, at folding temperature, Q_{CA} is higher than Q_{AA} for the folded ensemble. This suggests that although the backbone structure is native ($Q_{CA} \approx 0.8$), many of the native residue interactions form as temperature is decreased [Fig. 3(c,d)]. To account for this structurally and quantitatively, we calculated the difference between the probability of C_α contacts being formed $P(Q^i_{CA},Q_{CA})$ and AA contacts being formed $P(Q^i_{AA},Q_{CA})$ (Fig. 5). A value of 0 indicates that, on average, the C_α atoms of a residue pair are near their native distance when the side chains are in contact. Positive values are seen when extended side chains are interacting, resulting in the C_α atoms being far from their native distance. Negative values indicate backbone folding precedes side

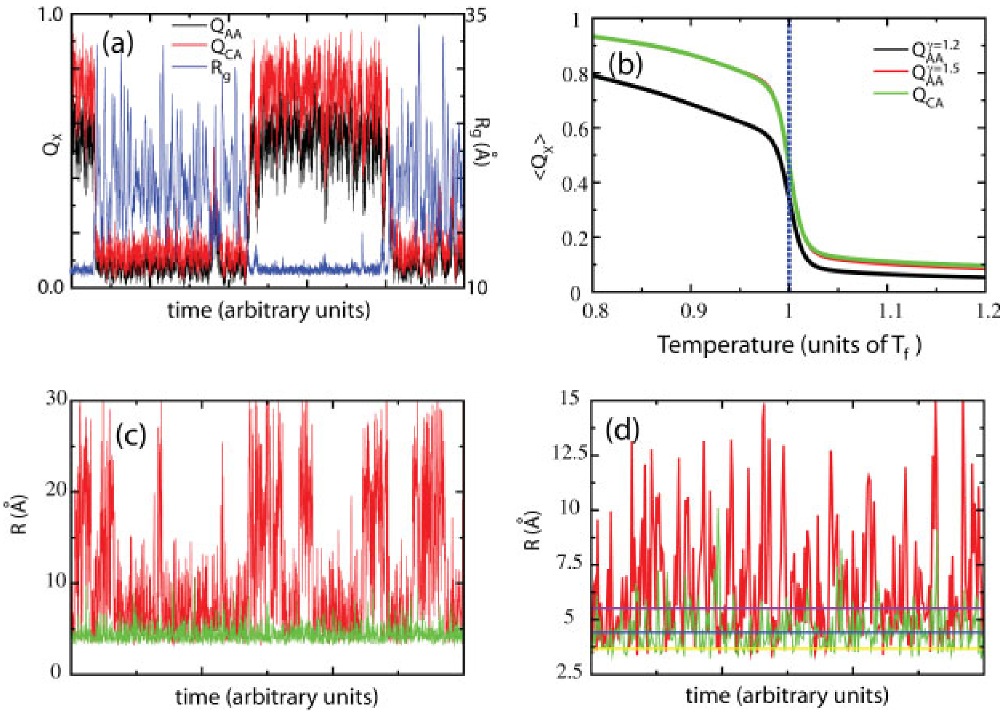

Figure 3

(**a**) Fraction of C_α contacts $Q_{CA}(t)$, AA contacts $Q_{AA}(t)$, and Radius of Gyration $R_g(t)$ as functions of time for a representative trajectory of CI2 with the AA model. (**b**) Average structure formation for several reaction coordinates. A contact between residues is formed when a single atom–atom contact between them is formed. An atom–atom contact is considered formed when the pair is at a distance $r < \gamma\sigma$ where σ is the native pair distance. The fraction of native residue contacts formed Q_{AA}^X is shown for $\gamma = 1.2$ (black) and $\gamma = 1.5$ (red). A C_α contact is formed when the C_α atoms are within 1.2 times their native distance (green). All three coordinates capture the same folding events. (**c**) Atom–atom distance for a contact in the active loop of CI2 versus time at T_f (red) and $T < T_f$ (green). Large changes in distance (>20 Å) coincide with folding transitions. Side chain rearrangements in the folded state ($R < 10$ Å) occur on much faster time scales than folding of the entire protein. (**d**) Same as Figure (**c**) with time scale decreased by a factor of 100. Horizontal lines correspond to σ (yellow), 1.2σ (blue), and 1.5σ (purple). As temperature is decreased, distance fluctuations and average distances decrease.

chain ordering.* Side chains in Protein A seem to be well-packed, in that there is concomitant side chain and backbone folding. In SH3, the turns have negative values, and are thus underpacked. In CI2, underpacking is primarily found in the active site loop and the C-terminal tail. These results reveal a signature of complicated folds:[52,53] a small subset of native contacts is sufficient to constrain the backbone to its native orientation, resulting in significantly underpacked regions in the native state. This occurs in complicated folds because an individual contact can impose a high level of order on the system. To form contacts that are distant in sequence, a large number of residues must also order. In Protein A, many contacts are local and only constrain single helical turns. In SH3 and CI2, fewer contacts are required to constrain the entire backbone (including the turns and loops).

Figures 3(c, d) shows the dynamics of a typical underpacked contact. As T is lowered below T_f the underpacked contact's average distance and distance fluctuations smoothly decrease. This results in a gradual increase in Q without a noticeable free-energy barrier [See Fig. 6(e)]. We hope that these subtle dynamics will be experimentally probed and tested in the future.

Understanding free-energy profiles through parametric variation: free-energy profiles can be altered through parametric changes

Although the folding mechanisms are stable, the free-energy barriers associated with folding and the locations of the folded basins vary systematically with parameters. Figure 6 shows free-energy profiles for SH3, CI2, and Protein A for several values of $R_{C/D}$ with $R_{BB/SC} = 2.0$. There are four distinct, interrelated, trends shared by all three proteins. First, there are two folding processes: backbone collapse and side chain packing. Second, the free energy minimum for the folded state moves to lower Q with increasing $R_{C/D}$. Third, the free energy barrier

*Q_{AA} is a generous definition of side chain packing, because a side chain is "packed" when one or more atom–atom contacts are formed. Thus, "underpacked" residues clearly have very little native structure.

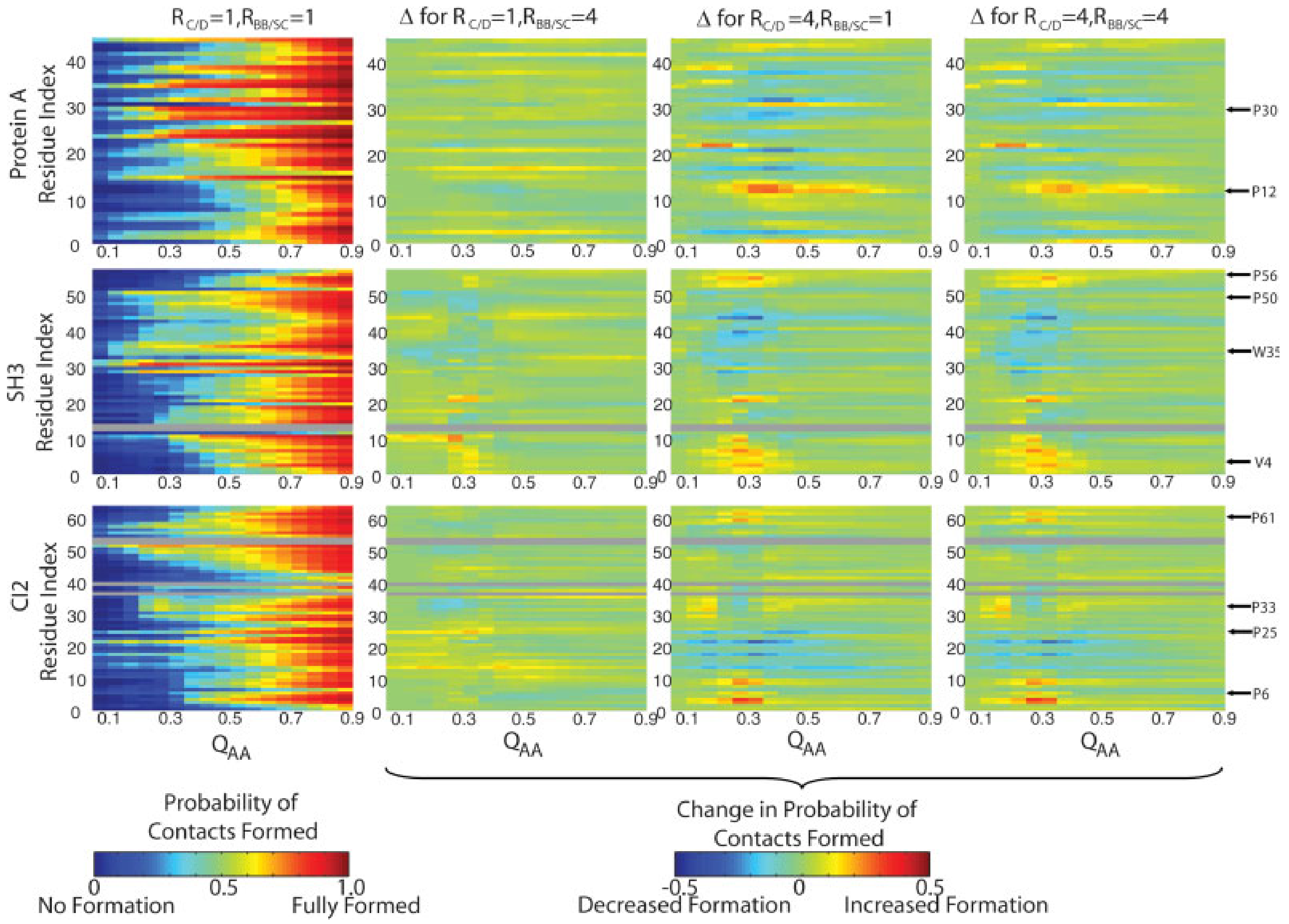

Figure 4

The left column shows the probability of contacts being formed for each residue $P(Q_i,Q_{AA})$ as a function of Q_{AA} for $R_{C/D}$ = 1.0 and $R_{BB/SC}$ = 1.0. The three right columns show $P(Q_i,Q_{AA})$ for different Hamiltonians relative to $R_{C/D}$ = 1.0 and $R_{BB/SC}$ = 1.0. Blue indicates a decrease in formation, relative to $R_{C/D}$ = 1.0 and $R_{BB/SC}$ = 1.0, and red an increase. Proline containing regions are often sensitive to contact energy. In Protein A, both P12 and P30 fold earlier with increased contact strength. In SH3, the increase in formation of Val4 may be attributed to interactions with Pro56, though Pro50 and Trp35 do not exhibit increased formation. In CI2, both Pro6 and Pro61 exhibit increased formation with increased contact strength. Residues that lack native contacts are shown in gray.

decreases with increasing $R_{C/D}$. Finally, increasing $R_{BB/SC}$ has similar effects as increasing $R_{C/D}$ (not shown).

The free-energy basins for the folded states are located at $Q_{CA} \approx 0.8$ and $Q_{AA} \approx 0.5$ [Fig. 6(d)], indicating that the backbone orders while many native atom–atom interactions remain extended. Thus, the entropy loss during the cooperative folding transition is likely dominated by backbone ordering. Side chain packing occurs both concomitantly with, and after, backbone ordering.

There are likely two major factors that lead to the observed trends. First, increasing $R_{C/D}$ increases contact strength. As seen in other simplified models,[54] when each contact is stronger, a smaller number of contacts is required (lower Q) to provide an equal amount of stabilizing energy. The second contributing factor is the change in side chain entropy. Although entropy loss in the backbone dominates the collapse transition, the gradual side chain packing can also lead to shifting basins. Increasing $R_{BB/SC}$ or $R_{C/D}$ reduces the strength of side chain dihedrals, resulting in more mobile unfolded side chains. Therefore, there is an increased entropy loss per side chain upon folding ΔS_{sc} when $R_{C/D}$ or $R_{BB/SC}$ is increased. Because side chains can pack independently of the collapse transition, when ΔS_{sc} increases, a fraction of the side chain interactions extend, while leaving the overall fold intact. Because the folded basin shifts to lower Q, the overall structure required to form a stable fold is reduced. A reduced barrier height naturally results when the folded basin is less ordered.

Free-energy barriers, in conjunction with diffusion constants, provide a direct connection to experimental folding rates.[55–57] We find that the relative barrier heights calculated using our AA model are similar to those from a C_α model [Fig. 6(f)]. The relative barrier

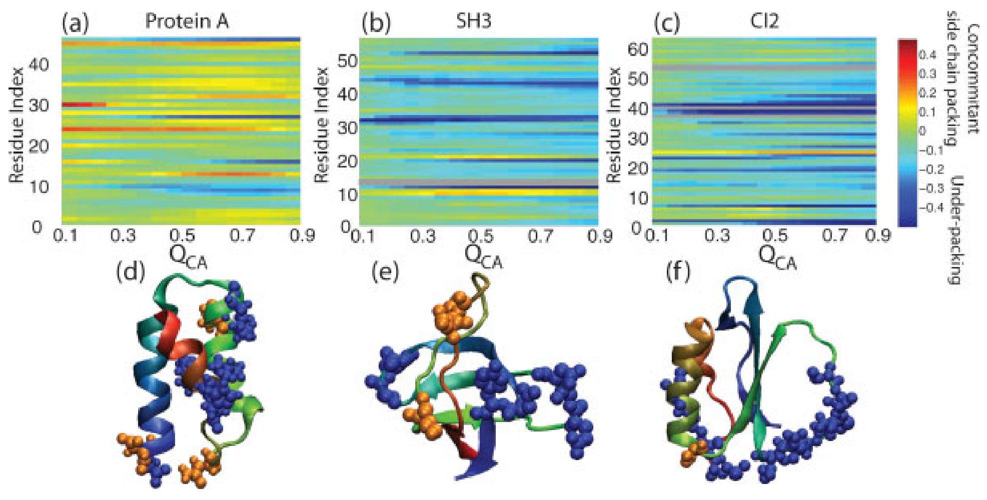

Figure 5

Difference in AA contact formation and C_α contact formation $P(Q^i_{AA},Q_{CA}) - P(Q^i_{CA},Q_{CA})$ for (**a**) Protein A, (**b**) SH3, and (**c**) CI2. Positive values (red) indicate that residues are interacting without the C_α atoms being near. Negative values (blue) indicate the residues are "underpacked": the C_α atoms are near each other without the side chains interacting. Residues that lack native contacts are shown in gray. (**d–f**) Underpacked (blue spheres) and well packed (orange spheres) residues are shown on the native structures. In Protein A, to order the backbone of a helix the side chains must be packed around it. Beta sheets are stabilized by nonlocal interactions. Thus, a small number of contacts can maintain the tertiary structure of SH3 without the side chains in the turn regions interacting, hence the underpacking. In CI2, the active site loop is significantly underpacked.

heights calculated from this model are known to correlate well with experimental rates.[57] We note in passing, that the absolute free-energy barriers in the AA model can be parametrically changed by up to a factor of two for a given protein and that the relative barrier heights between proteins remain constant. Thus, although the magnitude of the rates will be determined by the diffusion constant, the correlation between experimental

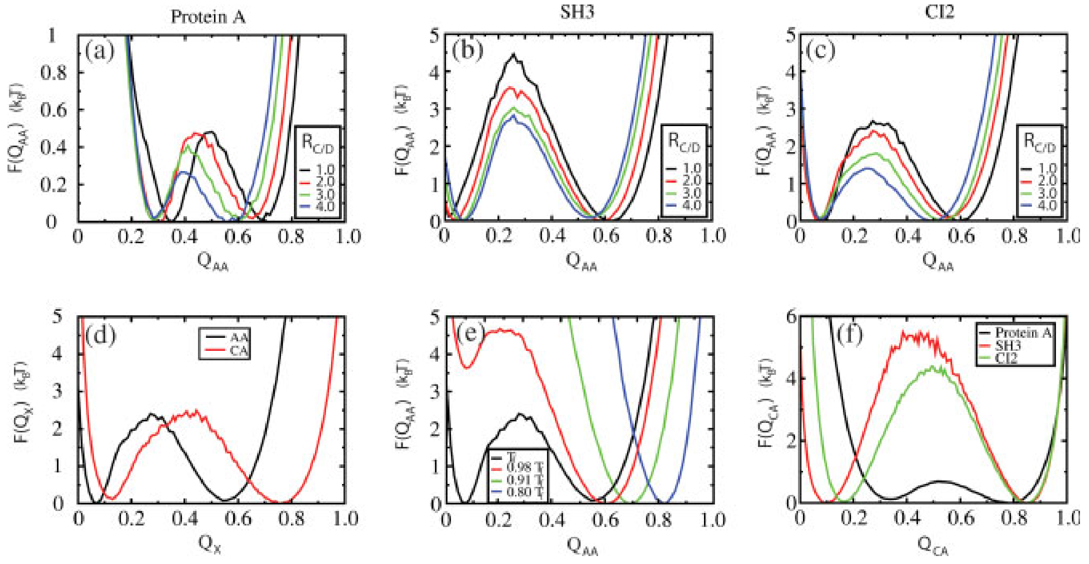

Figure 6

Free energy barriers in the AA model for (**a**) Protein A, (**b**) SH3, and (**c**) CI2. Profiles in (a–c) are for $R_{BB/SC} = 2.0$ with $R_{C/D} = 1.0$ (black), $R_{C/D} = 2.0$ (red), $R_{C/D} = 3.0$ (green), and $R_{C/D} = 4.0$ (blue). In SH3 and CI2, barrier height decreases and the folded basins move to lower Q with increasing $R_{C/D}$ and increasing $R_{BB/SC}$. (**d**) $F(Q_{CA}(t))$ and $F(Q_{AA}(T))$ for a typical parameter set demonstrate that the folded basins in (a–c) correspond to collapsed states. (**e**) Two distinct folding processes observed in our model: backbone collapse and side chain packing. (**f**) Free energy barriers obtained from C_α structure-based simulations for Protein A, SH3, and CI2. Barrier heights in the C_α simulations are greater than in AA simulations. Both models predict the largest barriers for SH3 and smallest for Protein A.

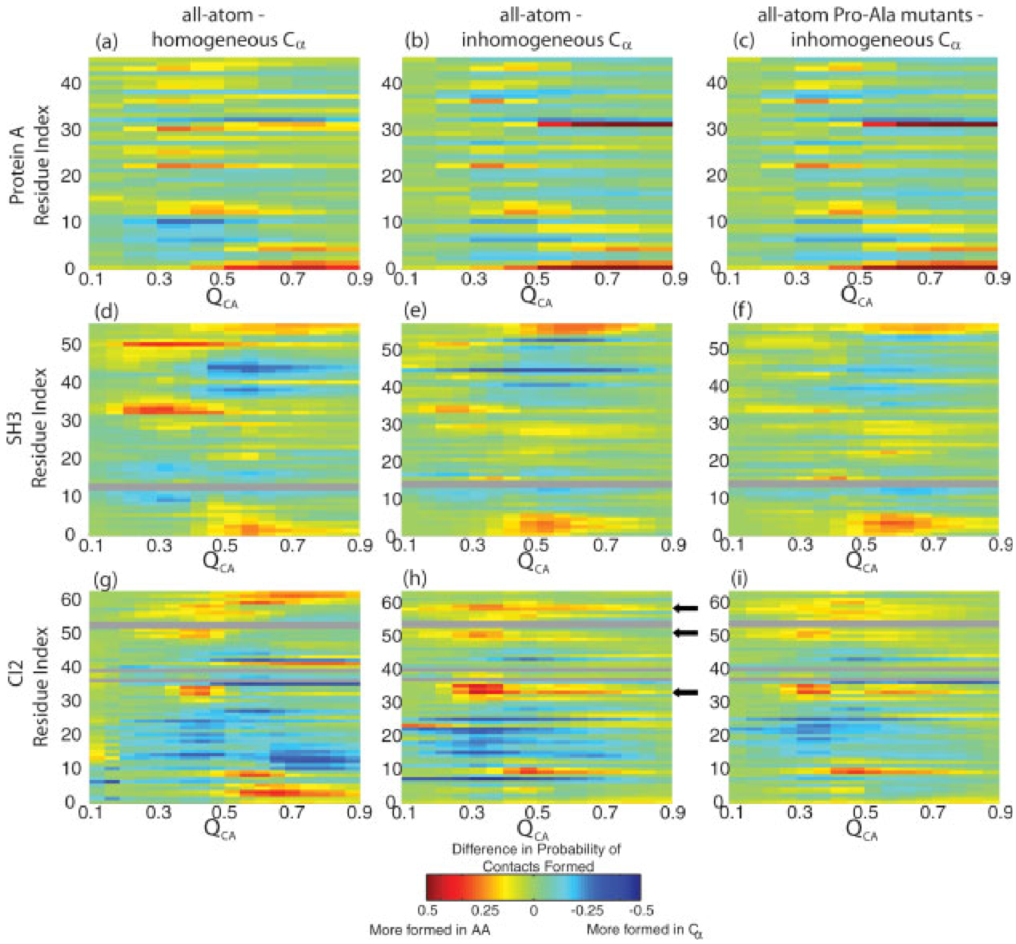

Figure 7

Comparison of backbone folding in C_α and AA structure-based models. The probability of contacts being formed in a C_α model, minus the probability of C_α contacts being formed in an AA model, is shown for (**a–c**) Protein A, (**d–f**) SH3, and (**g–i**) CI2. (a, d, g) Comparison of AA simulation to a C_α model with homogenous contact strength. (b, e, h) Comparison between AA results to an energetically inhomogeneous C_α model. Regions of increased formation in the AA representation correspond largely to proline containing regions, or regions that interact with proline, such as the minicore in CI2 (black arrows indicate mini-core residues), the tails of SH3 and turn 2 of Protein A. Increased formation in the tails of CI2 can largely be accounted for by the large number of contacts between GLU4 and ARG62. (c, f, i) The inhomogeneous C_α model compared to the AA model with all prolines mutated to alanines. Mutating proline to alanine improved agreement between models. Residues that lack native contacts are shown in gray.

folding rates and theoretical barriers is independent of the choice of parameters.

All-atom structure-based simulations capture C_α folding mechanism

Next, we compare the backbone folding mechanisms of our AA model and a commonly used C_α model.[9] The C_α representation has been successful at capturing experimentally determined protein folding mechanisms.[8,9,11] The first column in Figure 7 shows the differences in folding mechanisms between the AA model and an energetically homogeneous C_α model. Every contact and dihedral in the homogeneous C_α model has the same interaction strength. Because the AA model distributes contact energy inhomogeneously between residue pairs, it is not surprising that the mechanisms differ.

To remove differences arising from energetic homogeneity in the C_α model, we modified it such that each contact is weighted by the number of contacts between each residue pair in the AA model (Fig.7, second column). For Protein A this modification improves agreement. The remaining difference is in a single turn-to-tail contact [Gln1 with Ser31, Fig. 2(a)] that rarely forms in C_α simulations. In SH3, agreement improves around residues Asp34 and Asn52, while differences persist in Gln45 and the tails. The overall effect is increased formation around Gln45 at the expense of the tails. In CI2, there is significant agreement in the tails, though the mini-core still forms earlier (in the AA model), at the expense of

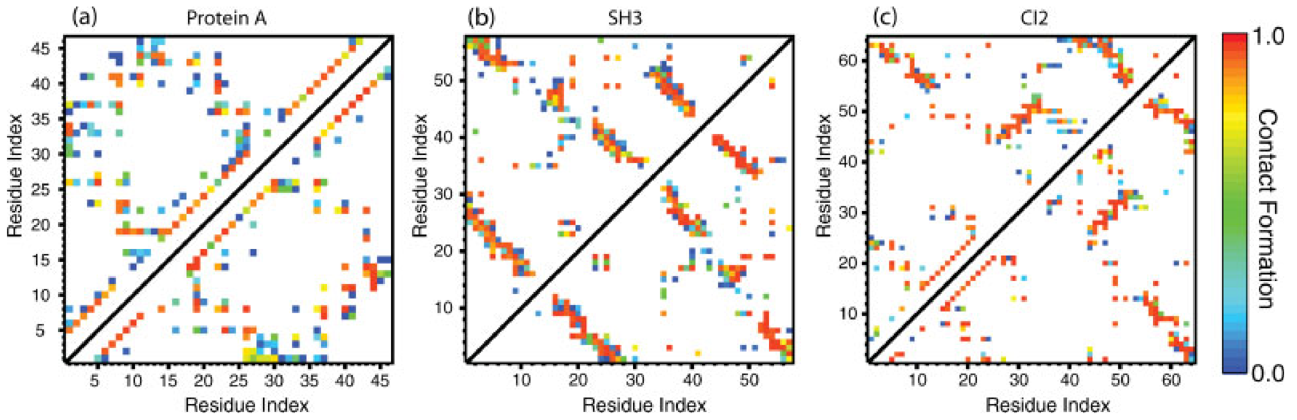

Figure 8

Probability of contacts being formed $P(i,j)$ at $T \approx 0.8\ T_f$ for the AA structure-based potential (top left) and an all-atom empirical forcefield (bottom right) for (**a**) Protein A, (**b**) SH3, and (**c**) CI2. Dark red indicates that residue i (x axis) and residue j (y axis) are always in contact under native conditions. Dark blue indicates the contact is formed rarely (less than 10% of the time). White indicates $P(i,j) < 0.025$. In all three proteins, contacts are more broadly distributed (higher number of low probability contacts) in the structure-based simulations than in all-atom empirical forcefield simulations (fewer contacts, but with higher probabilities). There are approximately four times as many contacts with $P(i,j) < 0.01$ for the structure-based simulations than are seen in all-atom empirical simulations, indicating more mobile dynamics.

the helix. For all three proteins, several regions of disagreement possess proline residues, whose $C_\delta—N$ bond is not included in the C_α model.

To eliminate effects specific to proline, we repeated the AA simulations with all prolines mutated to alanines. The third column of Figure 7 shows the Pro-Ala mutants compared to the inhomogeneous C_α model. Improved agreement is observed in Pro-Ala mutants of SH3 and CI2. In both proteins Pro-Ala mutations delay folding of proline regions, in agreement with proline effects on model stability. In SH3 the tails still form slightly earlier in the AA model, at the expense of residues 35–55. In CI2, the balance between minicore and helix formation is clearly improved, highlighting the importance of prolines in the folding process. Pro-Ala mutations have almost no effect on the folding mechanism of P12 and P30 in Protein A and P25 in CI2. This is likely because these prolines are located in turn regions. In our model, turns are highly constrained by short range contacts, and the reduced dihedral constraint (imposed by a proline) acts as a small perturbation. The remaining differences between the Pro-Ala AA mutants and the inhomogeneous C_α model demonstrate, to no surprise, that the inclusion of side chains alters the relative entropy of residues.

Native basin dynamics of AA structure-based model correlate with the dynamics of an all-atom empirical forcefield with explicit solvent

Two common measures of native state dynamics are native contact formation and root mean squared deviations in structure rmsd. Figure 8 shows the average contact formation in the native ensemble for the structure-based model and an all-atom empirical forcefield with an explicit solvent. Although the average contacts are not identical, no major differences in contact formation are observed. The overlaps between the AA maps and the all-atom empirical forcefield maps of Protein A, SH3, and CI2 are 0.85, 0.97, and 0.84, respectively. An overlap of 1 indicates identical maps, and 0 indicates the two maps have no contacts in common.

In a uniquely defined native state, the probability of each contact being formed is 1. Because we sample the native ensemble at finite temperatures, atom mobility leads to additional contacts being formed. In the structure-based model, these additional interactions are strictly repulsive. In an all-atom empirical forcefield these interactions can be attractive, yet they are observed more frequently in the structure-based model.[†] These contacts are likely due to increased mobility in the structure-based simulations. In all-atom empirical forcefields, hydration shells can result in less mobile side chains, and hence a narrower distribution of contacts.

The increased mobility is quantified by the structural rmsd. The magnitude of fluctuations in all-atom empirical simulations is much lower than in structure-based simulations (not shown). For the all-atom empirical forcefield at 300 K, the average rmsd for Protein A, SH3, and CI2 are 1.53, 1.00, and 0.97 Å. The rmsd of the C_α atoms are 1.23, 0.66, and 0.74. The same values are obtained in structure-based simulations at around $T = 0.55T_f$. In real temperature units, $0.55T_f$ corresponds to temperatures significantly less than 300 K. A likely cause

[†] In Figure 8 only interactions present more than 2.5% of the time are shown.

for the increased structural fluctuations is hydration effects of explicit solvent molecules in the all-atom empirical forcefield. To compare the distribution of rmsd fluctuations between models, correlation coefficients (r) were computed for the rmsd by atom in the all-atom empirical forcefield and the structure-based potential. For all parameter sets of the structure-based potential, $r \approx 0.7$ for CI2 and SH3 and $r \approx 0.8$ for Protein A.[‡]

DISCUSSION

In this article, we describe a systematic analysis of an AA structure-based model which bridges the gap between coarse-grained models and all-atom empirical forcefields. We show that in our C_α and AA structure-based models the global folding mechanisms agree and the main differences are largely due to energetic heterogeneity and the explicit representation of prolines in the AA model. Also, the native basin dynamics are similar in the AA structure-based model and an all-atom empirical forcefield with explicit solvent. In agreement with previous studies, the folding mechanisms in complicated folds are stable to parametric variation. On the other hand, the free-energy barriers associated with folding vary systematically with parameters. Because free-energy barriers are not a robust feature of this model, understanding the interplay between barrier heights and diffusion will be important before attempting to predict folding rates.[55,58,59]

Using this model we characterized two folding processes: one associated with backbone collapse and the other with side chain packing. We observed that backbone collapse is accompanied by partial side chain packing in a cooperative transition and residual side chain packing occurs as temperature is reduced below the global folding temperature. One explanation for the partial separation of backbone folding and side chain ordering may be that mobility in specific residues is necessary for the functional properties of proteins. Proteins are selected for their function. Orthogonal networks of residues responsible for stability and function have been proposed.[60,61] The observation in our model that some residues are not necessary to maintain the backbone structure is consistent with this proposal. In CI2, the backbone of the active site loop is in the native orientation, yet the side chains are not packed. In SH3, several turns are also disordered. Because binding sites are often found in loops, flexible loops may be more easily adapted to new sequences and functions.

Gradual side chain packing can also allow for proteins to functionally respond to cellular stress by affecting side chain orientations, without denaturing the entire protein. This is consistent with the prediction that localized unfolding, or cracking, is important for biological function of kinases and motor proteins.[15,18,62–67]

The current model explicitly includes the effects of topological contributions to protein folding, and the role of energetic contributions may now be elucidated. Our results are a significant step forward in understanding protein dynamics from the C_α to the all-atom level. In the coming years, it will be interesting to probe the effects of electrostatics, nonnative interactions, water, and explicit mutations in this model.

MODELS AND METHODS

Energy function

In our AA model of the protein, only heavy (non-hydrogen) atoms are included. Each atom is represented as a single bead of unit mass. Bond lengths, bond angles, improper dihedrals, and planar dihedrals are maintained by harmonic potentials. Nonbonded atom pairs that are in contact in the native state between residues i and j, where $i > j + 3$, are given a Lennard-Jones potential, whereas all other nonlocal interactions are repulsive. All contacts identified by the Contact of Structural Units software package (CSU)[68] were included. The functional form of the potential is,

$$
\begin{aligned}
V = \sum_{\text{bonds}} \varepsilon_r (r - r_o)^2 + \sum_{\text{angles}} \varepsilon_\theta (\theta - \theta_o)^2 \\
+ \sum_{\text{impropers/planar}} \varepsilon_\chi (\chi - \chi_o)^2 \\
+ \sum_{\text{backbone}} \varepsilon_{BB} F_D(\phi) + \sum_{\text{sidechains}} \varepsilon_{SC} F_D(\phi) \\
+ \sum_{\text{contacts}} \varepsilon_C \left[\left(\frac{\sigma_{ij}}{r}\right)^{12} - 2\left(\frac{\sigma_{ij}}{r}\right)^{6} \right] + \sum_{\text{non-contacts}} \varepsilon_{NC} \left(\frac{\sigma_{NC}}{r}\right)^{12} \qquad (1)
\end{aligned}
$$

where,

$$
F_D(\phi) = [1 - \cos(\phi - \phi_o)] + \frac{1}{2}[1 - \cos(3(\phi - \phi_o))] \qquad (2)
$$

and $\varepsilon_r = 100$, $\varepsilon_\theta = 20$, $\varepsilon_\chi = 10$, and $\varepsilon_{NC} = 0.01$. r_o, θ_o, χ_o, ϕ_o, and σ_{ij} are given the values found in the native state and $\sigma_{NC} = 2.5$ Å. When assigning dihedral strengths, we first group dihedral angles that share the middle two atoms. For example, in a protein backbone, one can define up to four dihedral angles that possess the same $C - C_\alpha$ covalent bond as the central bond. Each dihedral is given the interaction strength of $1/N_D$, where N_D is the number of dihedral angles in the group. ε_{BB} and ε_{SC} are then scaled so that $R_{BB/SC} = \frac{\varepsilon_{BB}}{\varepsilon_{SC}}$. Next, dihedral strengths and contact strengths are scaled such that our other system parameter, the ratio of total contact energy to total dihedral energy $R_{C/D} = \frac{\sum \varepsilon_C}{\sum \varepsilon_{BB} + \sum \varepsilon_{SC}}$, is satisfied. The total stabilizing energy is equal for all parameter sets (i. e. $\sum \varepsilon_C + \sum \varepsilon_{BB} + \sum \varepsilon_{SC} = \text{Constant}$).

As a reaction coordinate we use Q_{AA} and Q_{CA}. Q_{AA} is the fraction of natively interacting residues that are in

[‡]Comparison of rmsd of the C_α atoms yields similar values of r.

contact. Two residues are considered in contact if any native atom–atom interactions between the residues are within 1.2 times the native distance σ_{ij}. At $1.2\sigma_{ij}$ the potential energy of a native pair is approximately half of the minimum. Similarly, Q_{CA} is the fraction of natively interacting residue pairs whose C_α atoms are within 1.2 times their native distance.

Proline to alanine mutations

To investigate the role of proline residues in the AA model, proline to alanine mutants were constructed. This was achieved by removing the C_γ and C_δ atoms of each proline. Native contacts formed with the C_γ and C_δ of a proline were included as contacts with the C_β of the corresponding alanine. This ensured the energetics of the system were unperturbed, and only topology was modified.

Simulation details

All-atom structure-based simulations were performed using the GROMACS software package.[26] No modifications to the source code were necessary. Reduced units were used. The timestep τ was 0.0005. The Berendsen algorithm[69] was used[§] with the coupling constant of 1. For all folding results in this article several constant temperature runs were performed, with temperatures that corresponded to the protein being always folded to always unfolded. The Weighted Histogram Analysis Method[70,71] was used to combine data from multiple temperatures into single free-energy profiles.

All-atom empirical forcefield simulations

All-atom empirical forcefield simulations were performed using GROMACS,[26,72] with the OPLS-AA forcefield[73] with TIP3P water molecules.[74] Each protein was simulated for 10 ns at $T = 300$ K and a pressure of 1 atm. A timestep of 2 fs was used in conjunction with the LINCS[75,76] algorithm for constraining covalent bonds with hydrogen. Protein A, SH3, and CI2 were simulated with 2810, 3617, and 4644 water molecules in cubic boxes of initial dimensions 45.15Å, 48.98 Å, and 53.07 Å. Temperature was maintained using the Berendsen algorithm.[69] One nanosecond was allowed for equilibration. For the remaining 9 ns, structures were saved at 1 ps intervals.

Comparison of contacts

In the all-atom empirical forcefield simulations contacts were determined for each saved structure using CSU.[68] The average number of contacts $\langle Q \rangle$ was calculated for each protein. The probability of individual contacts being formed was averaged over all structures with $Q = \langle Q \rangle$. With the all-atom empirical potential $\langle Q \rangle$ was 80, 135, and 146 for Protein A, SH3, and CI2. This analysis was repeated for folded simulations with our AA structure-based simulations. For the structure-based simulations $\langle Q \rangle$ was 80, 138, and 144. To compare contact maps, the dot product of the two maps was taken.

[§] When using the Berendsen thermostat, numerical instabilities can arise when the bath-molecule coupling timescale is shorter than the timescale for internal energy diffusion. In our experience, these problems tend to surface when you simulate weakly interacting domains with implicit solvation. Because the present study investigates folding of single domain proteins under weak temperature coupling, these features are not likely a source of significant errors. Nonetheless, future work will also employ Langevin or Nose-hoover temperature coupling.

ACKNOWLEDGMENTS

We would like to thank Angel Garcia and Peter G. Wolynes for useful discussions regarding all-atom modeling.

REFERENCES

1. Leopold PE, Montal M, Onuchic JN. Protein folding funnels — a kinetic approach to the sequence structure relationship. Proc Natl Acad Sci USA 1992;18:8721–8725.
2. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Funnels, pathways, and the energy landscape of protein-folding–a synthesis. Proteins 1995;21:167–195.
3. Bryngelson JD, Wolynes PG. Spin-glasses and the statistical mechanics of protein folding. Proc Natl Acad Sci USA 1987;84:7524–7528.
4. Onuchic JN, Wolynes PG. Theory of protein folding. Curr Opin Struct Biol 2004;14:70–75.
5. Ueda Y, Taketomi H, Go N. Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effects of specific amino acid sequence represented by specific inter-unit interactions. Int J Pept Res 1975;7:445–459.
6. Shoemaker BA, Wang J, Wolynes PG. Structural correlations in protein folding funnels. Proc Natl Acad Sci USA 1997;94:777–782.
7. Nymeyer H, Garcia AE, Onuchic JN. Folding funnels and frustration in off-lattice minimalist protein landscapes. Proc Natl Acad Sci USA 1998;95:5921–5928.
8. Clementi C., Jennings PA, Onuchic JN. Prediction of folding mechanism for circular-permuted proteins. J Mol Biol 2001;311:879–890.
9. Clementi C, Nymeyer H, Onuchic JN. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. J Mol Biol 2000;298:937–953.
10. Gosavi S, Chavez LL, Jennings PA, Onuchic JN. Topological frustration and the folding of interleukin-1 beta. J Mol Biol 2006;357:986–996.
11. Levy Y, Onuchic JN. Mechanisms of protein assembly: lessons from minimalist models. Acc Chem Res 2006;39:135–142.
12. Levy Y, Cho SS, Shen T, Onuchic JN, Wolynes PG. Symmetry and frustration in protein energy landscapes: a near degeneracy resolves the Rop dimer-folding mystery. Proc Natl Acad Sci USA 2005;102:2373–2378.
13. Levy Y, Cho SS, Onuchic JN, Wolynes PG. A survey of flexible protein binding mechanisms and their transition states using native topology based energy landscapes. J Mol Biol 2005;346:1121–1145.
14. Yang SC, Cho SS, Levy Y, Cheung MS, Levine H, Wolynes PG, Onuchic JN. Domain swapping is a consequence of minimal frustration. Proc Natl Acad Sci USA 2004;101:13786–13791.
15. Whitford PC, Miyashita O, Levy Y, Onuchic JN. Conformational transitions of adenylate kinase: switching by cracking. J Mol Biol 2007;366:1661–1671.

16. Whitford PC, Gosavi S, Onuchic JN. Conformational transitions in adenylate kinase: allosteric communication reduces misligation. J Biol Chem 2008;283:2042–2048.
17. Schug A, Whitford PC, Levy Y, Onuchic JN. Mutations as trapdoors to two competing native conformations of the Rop-dimer. Proc Natl Acad Sci USA 2007;104:17674–17679.
18. Okazaki K, Koga N, Takada S, Onuchic JN, Wolynes PG. Multiple-basin energy landscapes for large amplitude conformational motions of proteins: Structure-based molecular dynamics simulations. Proc Natl Acad Sci USA 2006;103:11844–11849.
19. Best RB, Chen Y, Hummer G. Slow protein conformational dynamics from multiple experimental structures: the helix/sheet transition of Arc repressor. Structure 2005;13:1755–1763.
20. Zuckerman DM. Simulation of an ensemble of conformational transitions in a united-residue model of calmodulin. J Phys Chem B 2004;108:5127–5137.
21. Eastwood MP, Hardin C, Luthey-Schulten Z, Wolynes PG. Evaluating protein structure-prediction schemes using energy landscape theory. IBM J Res Dev 2001;5:475–497.
22. Rohl CA, Strauss CEM, Misura KMS, Baker D. Protein structure prediction using rosetta. Methods Enzymol 2004;383:66–93.
23. Harpaz Y, Elmasry N, Fersht AR, Henrick K. Direct observation of better hydration at the N terminus of an alpha-helix with glycine rather than alanine as the N-cap residue. Proc Natl Acad Sci USA 1994;91:311–315.
24. Humphrey W, Dalke A, Schulten K. VMD—visual molecular dynamics. J Mol Graph 1996;14:33–38.
25. Adcock SA, McCammon JA. Molecular dynamics: a survey of methods for simulating the activity of proteins. Chem Rev 2006;106: 1589–1615.
26. Lindahl E, Hess B, van der Spoel D. Gromacs 3.0: a package for molecular simulation and trajectory analysis. J Mol Mod 2001;7: 306–317.
27. Ponder JW, Case DA. Force fields for protein simulations. Adv Prot Chem 2003;66:27–85.
28. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, Schulten K. Scalable molecular dynamics with NAMD. J Comput Chem 2005;26:1781–1802.
29. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. Charmm: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 1983;4:187–217.
30. Eisenberg D, McLachlan AD. Solvation energy in protein folding and binding. Nature 1986;319:199–203.
31. Zhou R. Trp-cage: folding free energy landscape in explicit water. Proc Natl Acad Sci USA 2003;100:13280–13285.
32. Paschek D, Nymeyer H, Garcia A. Replica exchange simulation of reversible folding/unfolding of the Trp-cage miniprotein in explicit solvent: on the structure and possible role of internal water. J Struct Biol 2007;157:524–533.
33. Duan Y, Kollman PA. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science 1998;282:740–744.
34. Garcia A, Onuchic JN. FolDing a protein on a computer: an atomic description of the folding/unfolding of Protein A. Proc Natl Acad Sci USA 2003;100:13898–13903.
35. Schug A, Herges T, Wenzel W. Reproducible protein folding with the stochastic tunneling method. Phys Rev Lett 2003;91:158102.
36. Schug A, Verma A, Herges T, Lee KH, Wenzel W. Comparison of stochastic optimization methods for all-atom folding of the trp-cage protein. Chem Phys Chem 2005;6:2640–2646.
37. Schug A, Wenzel W. An evolutionary strategy for all-atom protein folding of the sixty-amino acid bacterial ribosomal protein l20. Biophys J 2006;90:4273–4280.
38. Jayachandran G, Vishal V, Pande VS. Using massively parallel simulation and Markovian models to study protein folding: examining the dynamics of the villin headpiece. J Chem Phys 2006;124:164902.
39. Freddolino PL, Liu F, Gruebele M, Schulten K. Ten-microsecond molecular dynamics simulation of a fast-folding WW domain. Biophys J 2008;94:L75–L77.
40. Clementi C, Garcia AE, Onuchic JN. Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: all-atom representation study of protein L. J Mol Biol 2003;326:933–954.
41. Shimada J, Kussell EL, Shakhnovich EI. The folding thermodynamics and kinetics of crambin using an all-atom monte carlo simulation. J Mol Biol 2001;308:79–95.
42. Linhananta A, Zhou Y. The role of sidechain packing and native contact interactions in folding: discontinuous molecular dynamics folding simulations of an all-atom Gō model of fragment B of Staphylococcal protein A. J Chem Phys 2002;117:8983–8995.
43. Zhou Y, Zhang C, Stell G, Wang J. Temperature dependence of the distribution of the first passage time: Results from discontinuous molecular dynamics simulations of an all-atom model of the second-hairpin fragment of protein G. J Amer Chem Soc 2003;125:6300–6305.
44. Linhananta A, Boer J, Mackay I. The equilibrium properties and folding kinetics of an all-atom Go model of the Trp-cage. J Chem Phys 2005;122:114901.
45. Gouda H, Torigoe H, Saito A, Sato M, Arata Y, Shimada I. Three-dimensional solution structure of the B domain of staphylococcal protein A: comparisons of the solution and crystal structures. Biochemistry 1992;31:9665–9672.
46. Xu W, Harrison SC, Eck MJ. Three-dimensional structure of the tyrosine kinase c-Src. Nature 1997;385:595–602.
47. Sato S, Religa TL, Daggett V, Fersht AR. Testing protein-folding simulations by experiment: B domain of protein A. Proc Natl Acad Sci USA 2004;101:6952–6956.
48. Viguera AR, Martinez JC, Filimonov VV, Mateo PL, Serrano L. Thermodynamic and kinetic-analysis of the sh3 domain of spectrin shows a 2-state folding transition. Biochemistry 1994;33:2142–2150.
49. Jackson SE, Fersht AR. Folding of Chymotrypsin Inhibitor 2. I. Evidence for a two-state transition. Biochemistry 1991;30:10428–10435.
50. Shea JE, Onuchic JN, Brooks CL, III. Probing the folding free energy landscape of the src-SH3 protein domain. Proc Natl Acad Sci USA 2002;99:16064–16068.
51. Hoang TX, Cieplek M. Sequencing of folding events in Go-type proteins. J Chem Phys 2000;113:8319–8328.
52. Shea JE, Onuchic JN, Brooks CL, III. Exploring the origins of topological frustration: design of minimally frustrated model of fragement B of protein A. Proc Natl Acad Sci USA 1999;96:12512–12517.
53. Plaxco KW, Simonsa KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 1998;277:985–994.
54. Prieto L, Rey A. Influence of the chain stiffness on the thermodynamics of a Gō-type model for protein folding. J Chem Phys 2007; 126:165103.
55. Bryngelson JD, Wolynes PG. Intermediates and barrier crossing in a random energy-model (with applications to protein folding). J Phys Chem 1989;93:6902–6912.
56. Kramers HA. Brownian motion in a field of force and the diffusion model of chemical reactions. Physica 1940;7:284–304.
57. Chavez LL, Onuchic JN, Clementi C. Quantifying the roughness on the free energy landscape: entropic bottlenecks and protein folding rates. J Am Chem Soc 2004;126:8426–8432.
58. Socci ND, Onuchic JN, Wolynes PG. Diffusive dynamics of the reaction coordinate for protien folding funnels. J Chem Phys 1996; 104:5860–5868.
59. Chahine J, Oliveira RJ, Leite VBP, Wang J. Configurational-dependent diffusion can shift the kinetic transition state and barrier height of protein folding. Proc Natl Acad Sci 2007;104:14646–14651.
60. Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R. Evolutionary information for specifying a protein fold. Nature 2005;437:512–518.

61. Gosavi S, Whitford PC, Jennings PA, Onuchic JN. Extracting function from a β-trefoil folding motif. Proc Natl Acad Sci USA 2008;105: 10384-10389.
62. Miyashita O, Onuchic JN, Wolynes PG. Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. Proc Natl Acad Sci USA 2003;100:12570–12575.
63. Miyashita O, Wolynes PG, Onuchic JN. Simple energy landscape model for kinetics of functional transitions in proteins. J Phys Chem B 2005;109:1959–1969.
64. Whitford PC, Onuchic JN, Wolynes PG. The energy landscape along an enzymatic reaction trajectory: hinges or cracks? HFSP J 2008;2:61–64.
65. Hyeon C, Onuchic JN. Internal strain regulates the nucleotide binding site of the kinesin leading head. Proc Natl Acad Sci USA 2007;104:2175–2180.
66. Hyeon C, Onuchic JN. Mechanical control of the directional stepping dynamics of the kinesin motor. Proc. Natl Acad Sci USA 2007;104:17382–17387.
67. Wonmuk H, Lang MJ, Karplus M. Force generation in kinesin hinges on cover-neck bundle formation. Structure 2008;16:62–71.
68. Sobolev V, Wade R, Vried G, Edelman M. Molecular docking using surface complementarity. Proteins: Struct Funct Genet 1996;25:120–129.
69. Berendsen HJC, Postma JPM, VanGunsteren WF, Dinola A, Haak JR. Molecular-dynamics with coupling to an external bath. J Chem Phys 1984;81:3684–3690.
70. Ferrenberg AM, Swendsen RH. New Monte Carlo technique for studying phase transitions. Phys Rev Lett 1988;61:2635–2638.
71. Ferrenberg AM, Swendsen RH. Optimized Monte-Carlo data analysis. Phys Rev Lett 1989;63:1195–1198.
72. Berendsen HJC, van der Spoel D, van Drunen R. GROMACS: a message-passing parallel molecular dynamics implementation. Comput Phys Comm 1995;91:43–56.
73. Jorgensen WL, Tirado-Rives J. The OPLS potential functions for proteins. energy minimizations for crystals of cyclic peptides and crambin. J Am Chem Soc 1988;110:1657–1666.
74. Jorgensen WL, Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. J Chem Phys 1983;79:926–935.
75. Miyamoto S, Kollman PA. SETTLE: an analytical version of the SHAKE and RATTLE algorithms for rigid water models. J Comput Chem 1992;13:952–962.
76. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: a Linear constraint solver for molecular simulations. J Comput Chem 1997; 18:1463–1472.