Distinguishing Foldable Proteins From Nonfolders: When and How Do They Differ?

Tobin R. Sosnick,[1]* R. Stephen Berry,[2] Andrés Colubri,[3] and Ariel Fernández[2,3]*
[1]*Department of Biochemistry and Molecular Biology and the Institute for Biophysical Dynamics, University of Chicago, Chicago, Illinois*
[2]*Department of Chemistry and the James Frank Institute, University of Chicago, Chicago, Illinois*
[3]*Instituto de Matemática, Universidad Nacional del Sur, Consejo Nacional de Investigaciones Científicas y Técnicas, Bahía Blanca, Argentina*

ABSTRACT **When a denatured polypeptide is put into refolding conditions, it undergoes conformational changes on a variety of times scales. We set out here to distinguish the fast events that promote productive folding from other processes that may be generic to any non-folding polypeptide. We have apply an ab initio folding algorithm to model the folding of various proteins and their compositionally identical, random-sequence analogues. In the earliest stages, proteins and their scrambled-sequence counterparts undergo indistinguishable reductions in the extent to which they explore conformation space. For both polypeptides, an early contraction occurs but does not involve the formation of a distinct intermediate. Following this phase, however, the naturally-occurring sequences are distinguished by an increase in the formation of three-body correlations wherein a hydrophobic group desolvates and protects an intra-molecular hydrogen bond. These correlations are manifested in a mild but measurable reduction of the accessible configuration space beyond that of the random-sequence peptides, and portend the folding to the native structure. Hence, early events reflect a generic response of the denatured ensemble to a change in solvent condition, but the wild-type sequence develops additional correlations as its structure evolves that can reveal the protein's foldability. Proteins 2002;49:15–23.** © 2002 Wiley-Liss, Inc.

Key words: protein folding; burst phase; transition state; folding pathway; nucleation; folding intermediate; Ramachandran basin; Shannon entropy

INTRODUCTION

The origin of submillisecond signals that are seen when a protein in a denaturing environment (good solvent) is mixed into a folding condition (poor solvent) is a subject of debate. One view is that these changes reflect the formation of a productive folding intermediate, I, with distinct native-like structure [$U \rightarrow I \rightarrow N$; Fig. 1(A)]. More simply, the changes could represent a readjustment of the denatured state ensemble, U, to the new solvent condition upon the dilution of denaturant [$U' \rightarrow U \rightarrow N$, Fig. 1(B), e.g., $U_{high} \rightarrow U_{low}$]. Such a readjustment would appear to be "obligatory and on-pathway" but would do little that we could characterize as steps along the conformation-space search for the native structure.

The "burst-phase" controversy is relevant to the folding of certain proteins including ubiquitin,[1,2] Rnase A,[3,4] and equine cytochrome c.[5–12] It should be appreciated, however, that the vast majority of proteins composed of fewer than 110 amino acids fold without forming any stable intermediates prior to the major folding event and lack such a burst phase.[2,13] Here, the term "intermediate" is used to describe species with some secondary and tertiary structure. Such species are to be distinguished from conformations containing some regions of stable local turn or helical structure, which may best be described as residual structure. These conformations are considered to be part of the denatured ensemble.

In order to understand the fast events, we examined folding pathways using a recently developed algorithm that is capable of largely reproducing the native contact map and secondary structures of various small proteins.[14–16] In this algorithm, the chain's conformation is coarsely specified for each residue by the occupation of discrete regions, or basins, in the Ramachandran Φ,Ψ plot. On folding trajectories, each residue moves stochastically from one basin to another at a rate that reflects the extent of its structural engagement, similar to the operational tenets of the LINUS routine.[17] This simplification enables folding trajectories to be followed from 1 ns to 10 msec or more, a time-scale that is computationally inaccessible by molecular dynamics.

Abbreviations: cyt c, cytochrome c; ΔE_i, virtual energy loss; Fl, fluorescence; h-h, hydrophobic-hydrophobic; h-p, hydrophobic-polar; LTM, Local Topology Matrix; σ, Shannon entropy; Ub, ubiquitin.

Grant sponsor: National Institutes of Health; Grant sponsor: National Science Foundation; Grant sponsor: Packard Foundation Interdisciplinary Science Program.

*Correspondence to: Tobin R. Sosnick, Department of Biochemistry and Molecular Biology and the Institute for Biophysical Dynamics, University of Chicago, Chicago, IL 60637. E-mail: trsosnic@midway.uchicago.edu or Ariel Fernández, Institute for Biophysical Dynamics University of Chicago, Chicago, IL 60637. E-mail: ariel@uchicago.edu

Received 4 February 2001; Accepted 22 April 2002

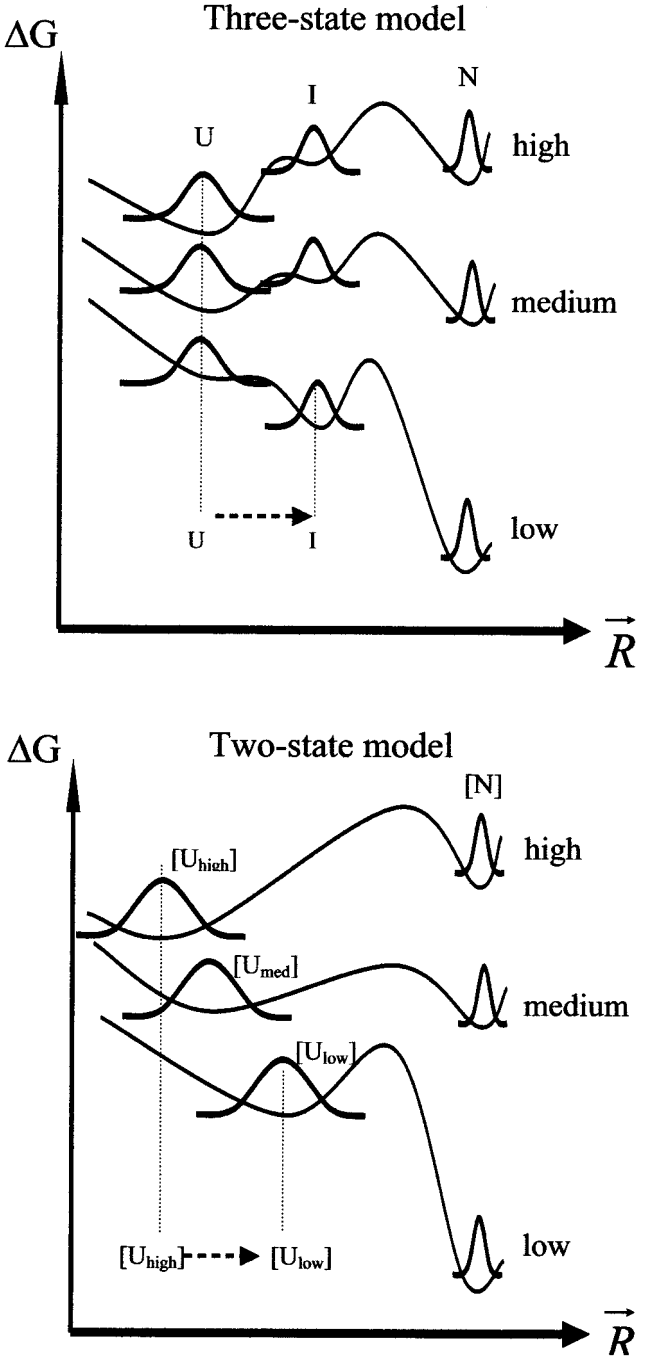

Fig. 1. Alternative explanations for the fast events in protein folding. Schematic free energy vs. configuration diagrams as a function of denaturant concentration for three-state folding pathway (**top**) with a structured intermediate that accumulates at lower denaturant concentrations, U→I→N; and a two-state folding pathway (**bottom**) where the denatured ensemble changes with denaturant level, $[U_{high}] \rightarrow [U_{low}] \rightarrow [N]$. The Gaussian-like curves illustrate the conformational diversity of a given ensemble. The rapid response to a change denaturant concentration from a high to low denaturant is illustrated by the dashed arrow. In the three-state model, the initial event represents a barrier-crossing process from U→I, whereas in the two-state model, the fast events reflect a barrier-free relaxation of the denaturant-dependent unfolded ensemble, $[U_{high}] \rightarrow [U_{low}]$.

The algorithm has been substantiated in its prediction of the folding pathway of β-lactoglobulin[15] found shortly afterward by experiment.[18] For the major kinetic intermediate, the following structural elements were proposed: β-strands D, F, G, and H, a native helical region (residues 136–142), and a non-native helical region (residues 14–50). Experimentally, Kuwata et al. observed protection from hydrogen exchange in β-strands F, G, H, and the adjacent native helix (129–141). Importantly, they observed a non-native helical region (residues 12–21), and based upon circular dichroism studies,[19] they concluded that an additional 16 residues from regions such as β-strands B-E (residues 40–80) also form non-native helical structure.

On the basis of this comparison, and the success in other studies,[14,16] we believe that the simulation does contain structural detail that is key to describing successful folding pathways of a protein. The chain's conformation is that of a true polypeptide, whose residues move with a Φ,Ψ dihedral angle rotation from one allowed Ramachandran basin to another. Furthermore, the simulation uses a non-bonding potential containing angular-dependent backbone-backbone hydrogen bonding, hydrophobic, and electrostatic interactions that differentiate the folding of a true polypeptide from that of a generic heteropolymer. These features of the model make it an approach toward identifying economic folding pathways. In this respect, it provides an approach different and separate from much of the current effort in protein folding studies.

Using this algorithm, we have investigated the interplay between different chain-condensation events at the earliest phases, so that the ability to fold a protein to its native structure is not a central inference of this study. We applied the algorithm to two small proteins that fold without the accumulation of intermediates. These are the 78-amino acid α/β protein mammalian ubiquitin[2,16,20] and the 80-amino acid helical protein λ-repressor.[21] We also studied a larger protein that folds with at least one distinct kinetic intermediate,[18,22] the 162-amino acid β-lactoglobulin. The folding trajectories are detailed for these proteins, and for several random-sequence counterparts of the two smaller proteins. By comparing their folding trajectories, we can distinguish the generic polypeptide events from the important folding events that lead productively to the native state, primarily on the basis of the latter's formation, after the burst phase, of protective three-residue contacts at the level of tertiary structure.

MATERIALS AND METHODS

Ab Initio Ramachandran Basin Folding Algorithm.

We believe that a meaningful understanding of the folding pathway can best be achieved by a compromise between a computationally-efficient lattice model with binary patterning and a computationally excessive all-atom model with explicit solvent. A natural, coarsened simplification is to specify each residue's backbone dihedral Φ,Ψ angles in terms of the four major allowable regions or basins in the Ramachandran plot. The first basin is compatible with extended or β-strand structures (basin no. 1), the second with α-helical (no. 2), and the

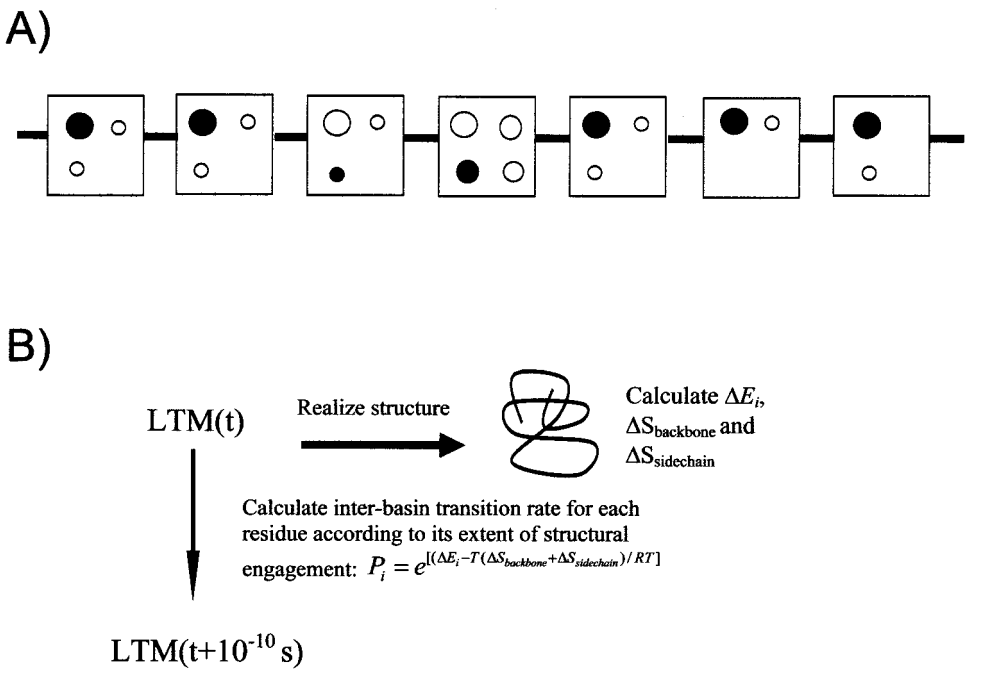

Fig. 2. Ab initio folding algorithm. **A:** LTM description of an illustrative 7-amino-acid chain composed of three alanine-like residues, a glycine, an alanine-like, a residue preceding a proline, and a proline to illustrate the four general classes of residues. The square represents the Ramachandran Φ,Ψ plane for each residue with the circles representing each of the allowable basins such as those that include extended (upper left, basin, no. 1), α-helical or turn (lower left, basin, no. 2) or 3_{10} helical conformations (upper right, basin, no. 3). The filled circles denote the conformation of the residue at a particular time, which for the 1,1,2,2,1,1,1 pattern shown, identifies the structure as a β-turn motif with a carboxy terminal proline. **B:** Flow chart of the algorithm.

third, with 3_{10}-helical (no. 3) structures; the more extended basin (no. 4) is available only to glycines. In the intermediate-resolution representation used here, the conformation of a polypeptide is defined by a series of 1, 2, 3, and 4's where each entry in the string represents the basin location for a residue.

This string, termed the Local Topology Matrix or LTM(t), is a coarse representation, but one which satisfies the inherent geometrical constraints of a real polypeptide chain [Fig. 2(A)]. The precise coordinates of the chain (i.e., the physical realization of a LTM) are defined by explicit Φ,Ψ angles. To maintain structural continuity during a folding trajectory, the explicit dihedral angles are retained for each residue from one time step to the next until that residue's basin is explicitly changed, as determined by the criteria noted below. The method is described in full detail in recent publications.[15,16,23–25]

The initial LTM is obtained by a random assignment of Ramachandran basins for individual residues with probabilities proportional to each basin's area. Likewise, the probability of subsequently moving to a particular basin is given by its relative lake area (microcanonical entropy). Residues can be classified generally as L-alanyl-like, glycine, proline, and residues preceding prolines (which have additional constraints). The areas of the basins reflect the observed distribution of Φ,Ψ angles in folded proteins for each type of amino acid. Each amino acid is further characterized by its chemical character, for example, hydrophobic, polar, neutral, or charged.

The rate of transition for a given residue depends upon the amount of structure constraining it (Fig. 2B). Explicitly, every 10^{-10} s, the probability of the i^{th} residue changing basins is given by $P_i = e^{[(\Delta E_i - T(\Delta S_{backbone} + \Delta S_{sidechain})/RT]}$. The virtual energy loss, ΔE_i, reflects the dismantling of structure if the i^{th} residue changes basin; $\Delta S_{backbone}$ and $\Delta S_{sidechain}$ are the total backbone and side-chain configurational entropies released by the change of basin (see below). The energetic term, ΔE_i is calculated from the protein's structure, assuming that a (virtual) change in the i^{th} residue's dihedral angle causes the loss of the interactions between j,k-pairs of residues located on flanking sides of the i^{th} residue ($j < i < k$). In effect, the exponent, $\Delta E_i - T(\Delta S_{backbone} + \Delta S_{sidechain})$, gives the extent of structural engagement of residue i in thermodynamic terms.

Upon basin transition, the initial dihedral angle is obtained from an intra-basin distribution given by the program PROCHECK.[26] Subsequently, the angle is varied within the new basin to reduce non-bonded energy using a Monte-Carlo optimization subroutine.[16]

About half of the folding trajectories largely reproduced the native Ramachandran basins (i.e., native LTM) and contact map.[16] The remaining runs became irreversibly trapped in a cluster of misfolded structures at long times and were discarded.

Potential Energy Function

The importance of local environment on pair-wise interactions is increasingly evident.[27–32] We have recently

found that making hydrogen bonds context-sensitive results in more expeditious and robust folding behavior.[16] Solvent ordering around hydrophobic side chains inhibits backbone solvation and enhances intra-protein hydrogen bonding between two other residues. This situation is analogous to the proposal that low concentrations of 2,2,2-trifluoroethanol (TFE) act by promoting desolvation of the polypeptide backbone, instead of directly strengthening hydrogen bonds.[33]

Our method requires an effective non-bonded potential that must achieve two goals: (1) accurately represent the different intra-molecular interactions and (2) incorporate implicitly the solvent effect on residue-residue interactions. The solvent effect was operationally incorporated in the intramolecular potential by introducing an effective hydrophobic attractive contribution, a repulsive hydrophobic-polar two-body contribution, and a rescaling of all pair-wise contributions according to the desolvation levels of the interacting residues. In view of this, our effective potential should be regarded as representing the enthalpic contributions rather than energetic contributions associated with intramolecular interactions.

We model the long-range non-bonded interactions between the residues by including the following terms in our effective potential:

$$U_{\mathrm{nb}} = U_{\mathrm{LJ}} + U_{\mathrm{solv}} + U_{\mathrm{coul}} + U_{\mathrm{dip}} + U_{\mathrm{Hbond}} \tag{1}$$

where U_{LJ} represents a Lennard-Jones contribution that accounts for excluded volume, U_{solv} is the effective solvophobic term accounting for the attraction between hydrophobic residues and the repulsion between hydrophobic and polar residues, U_{coul} represents the ionic energy between charged side-chains, U_{dip} models the backbone dipole-dipole interactions, and U_{Hbond} corresponds to the backbone hydrogen bonding.

In a zeroth-order approximation, generically denoted U^0, each of these terms can be expressed as a sum over pairwise contributions: $U^0 = \sum_{ij} U^0(i, j)$. Under this approximation, the potential function does not reflect the effect of local solvent environments on the stability of the dielectric-dependent interactions (solvophobic, coulombic, and hydrogen-bond). In order to incorporate this effect, we rescale the zeroth-order contribution for each (i, j) pair, $U^0(i, j)$, by introducing renormalization factors f_i and f_j, which depend on the level of desolvation of residues i and j. Thus the rescaled pairwise energy is $U(i, j) = f_i f_j U^0(i, j)$, where $f_i = f_i(L_i)$ and L_i = extent of desolvation of residue i. For each residue i, we define the variable $L_i = V_H(i)/V_T$, where V_T is the total volume of neighborhood-defining ball of radius 7 Å, centered at the α-carbon i, and $V_H(i)$ is the volume of this sphere taken up by hydrophobic side-chains that are close enough to i to be included in its neighbor-defining sphere. From this definition, it follows that L_i has a value between 0 and 1, which reflects the amount of hydrophobic burial of the residue. Under this ansatz, $L_i \sim 0$ means that the residue *i* is completely exposed to the solvent, while $L_i \sim 1$ means that residue *i* is totally buried.

The functional dependence of the factors f_i on the desolvation level L_i models quantitatively how the solvent affects the effective stability of dielectric-dependent interactions. For example, desolvating a hydrogen bond makes it effectively stronger with respect to a solvent-exposed bond. On the other hand, the hydrophobic attraction engaging a desolvated pair becomes effectively weaker in accord with the relative burial of solvent-exposed surface: the entropy-driven hydrophobic effect is nonexistent in the absence of surrounding solvent.

The effective solvophobic potential can be expressed as the sum of an attractive and a repulsive term**:** $U_{solv} = U_{solv,\,hh}, + U_{solv\,hp}$. The solvophobic hydrophobic-hydrophobic (h-h) attraction is associated with the minimal ordering of solvent around nonpolar moieties. The hydrophobic-polar (h-p) proximity between a polar and a nonpolar side-chain is energetically unfavorable because the solvent shells generated around each residue are incompatible with each other due the different relative orientation of the solvent dipoles.

The pairwise contributions of both attractive and repulsive terms are modeled as

$$U^0_{\mathrm{solv},\,hh}(i,j) = c_{hh}(i)\, c_{hh}(j)\, F_{hh}(r_{ij}) \tag{2a}$$

$$U^0_{\mathrm{solv},\,hp}(i,j) = c_{hp}(i)\, c_{hp}(j)\, F_{hp}(r_{ij}) \tag{2b}$$

where r_{ij} is the distance between the α-carbons *i* and *j*. In the case of the solvophobic attraction, $F_{hh}(r_{ij})$ represents a potential well with a depth of -3.1 kcal mol^{-1}, which ranges from 4.5 to 6.7 Å. The value of $c_{hh}(i)$ is proportional to the effective surface area of the side-chain of residue *i*. For the solvophobic repulsion, $F_{hp}(r_{ij})$ represents a potential bump with a height of 1 kcal mol^{-1}, over the range from 3.5 to 6.7 Å. For the h-p repulsion with *i, j* being hydrophobic and polar, respectively, then $c_{hp}(i)=c_{hh}(i)$, while $c_{hp}(j)$ is proportional to the polarity of residue j. The ranges for both terms were adjusted by inspecting the typical distances between the α-carbons in native structures.

The single-residue renormalization factor for the solvophobic potential was constructed to decrease monotonically from 1 to 0 upon the increase from 0.7 to 1 in the desolvation parameter L_i, reflecting the fact that both the solvophobic attraction and the solvophobic repulsion become progressively weaker as the amount of surrounding solvent decreases. These parameters are consistent with typical survival timescales of structures with different levels of solvent exposure.[14–16]

The effective ionic interaction between two charged residues in vacuum (zeroth-order) is represented by the function $U^0_{coul}(i, j) = q(i)\, q(j)\, F_{coul}(r_{ij})$, where r_{ij} is the distance between the α-carbons i and j, q(i) and q(j) are the effective dimensionless charges of the amino acids and $F_{coul}(r_{ij}) = c_{coul}/ r_{ij}$. c_{coul} is an adjustable parameter that was fixed at 7.5 kcal mol^{-1} Å^{-1}. An effective charge is assigned to the amino acids at neutral pH without any renormalization due to conformation-related environments.[16]

We rescale the ionic interaction depending on the desolvation levels of the residues, as the ionic interaction depends on the effective dielectric of the surrounding

medium. Because the solvent molecules tend to screen the charges of the side-chains, we use a renormalization factor that grows from 0.5 to 1.2 as the desolvation level of L_i ranges from 0 to 0.8. The adopted values are in accord with compiled data on context-dependent salt bridge stabilities and burial of coulombic interactions.[16,25,34]

Our zeroth-order hydrogen-bond energy between residues *i* and *j* is dependent on the relative orientation of the intervening residues and has the following generic form:

$$U^0_{\mathrm{Hbond}}(i,j) = E_{\mathrm{Hbond}}(H_i, O_j) + E_{\mathrm{Hbond}}(H_j, O_i) \qquad (3)$$

that accounts for the two possible backbone hydrogen bonds that can be formed between the residues i and j. The function $E_{\mathrm{Hbond}}(H_i, O_j)$ is the following product

$$E_{\mathrm{Hbond}}(H_i, O_j) = G_{\mathrm{Hbond}}(H_i, O_j) \cdot F_{\mathrm{Hbond}}(H_i, O_j) \qquad (4)$$

where the factor $G_{\mathrm{Hbond}}(H_i, O_j)$ penalizes the departure of the hydrogen bond from linearity, and $F_{\mathrm{Hbond}}(H_i, O_j)$ keeps the H_i-O_j distance between 1 and 3 Å. For a geometrically perfect hydrogen bond, we have $E_{\mathrm{Hbond}}(H_i, O_j) = -1.1$ kcal mol^{-1}.

The ability of water to displace intramolecular hydrogen bonds is well established. Thus, intramolecular hydrogen bond survival is contingent on the chain finding a protective topology that inhibits the distortion of surrounding solvent structure, thereby preventing water attack. Based on this fact, the renormalization factor used to rescale the effective energy of a hydrogen bond has been assumed to increase linearly from 0 to 2 as the desolvation level, L_i, goes from 0 to 0.7, and remains constant until the desolvation level reaches 1.

The loss of backbone entropy for each residue is estimated from the relative area (microcanonical entropy) of the Ramachandran basin it occupies. The side-chain entropy loss of a flexible chain is reduced commensurately with the number of contacts it forms and is proportional to the number of side-chain torsional modes.

RESULTS AND DISCUSSION

Folding of Ubiquitin

In earlier studies, the algorithm was applied (in a somewhat more primitive form) to ubiquitin (Ub). Multiple folding trajectories resulted in the native Ramachandran basins for each residue (i.e., the LTM) as well as the contact map.[16] For a single successful trajectory, the number of residues changing basins per unit time, I(t) indicates that many residues are hopping from basin to basin up to a critical time, t*=48 msec [Fig. 3(A)]. After this point, fluctuations are completely quenched and the native structure persists.

To assess whether the early-time fluctuations are localized in certain regions while other parts form persistent structure (the signature of an on-pathway intermediate), the fluctuations are examined for a small set of consecutive residues across the whole sequence [Fig. 3(B)]. This figure is a new kind of representation of the evolving structure of the protein backbone, showing the history of basin occupancy. In Figure 3(B), the basin location averaged over a three-residue window indicates that all regions undergo extensive backbone fluctuations. The correlations do not persist longer than 2.84 msec, and hence do not represent the formation of a distinct species. The effects of long-lived correlations can be detected only in later contact maps near the transition state at 48 msec [Fig. 3(B), Fig. 4], for example, where one correlation is seen to stabilize the amino-terminus of the α-helix in Ub.

In other words, native and non-native structures, both local and non-local, transiently form and dismantle, until

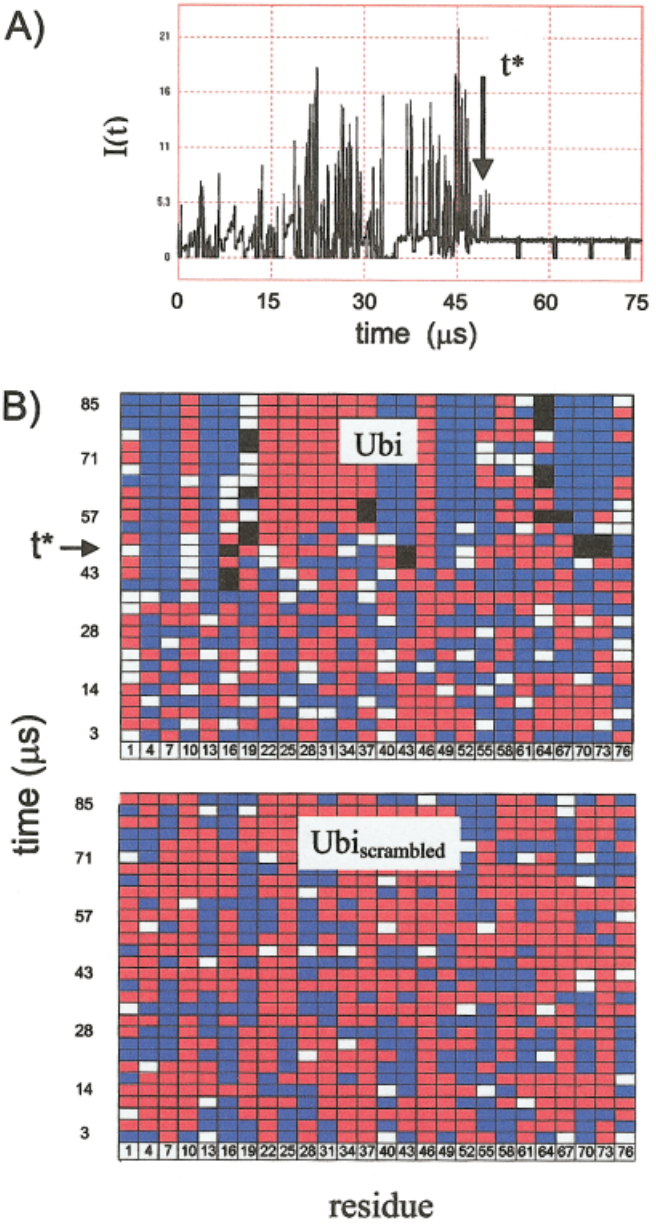

Fig. 3. Properties of an individual folding trajectory. **A:** Number of basins changing per unit time, I(t), for Ub. The nucleation event at t* = 4.8×10^{-5}s is noted. **B:** Time evolution of the basin location of groupings of three residues, Ub (**top**) and $\mathrm{Ub}_{\mathrm{scrambled}}$ (**bottom**), averaged every 2.84 msec and colored according to basin location: extended or β sheet (blue), α helix or turn (red), and 3_{10} helix (white) regions. A black square represents a hydrophobic residue involved in a three-body correlation.

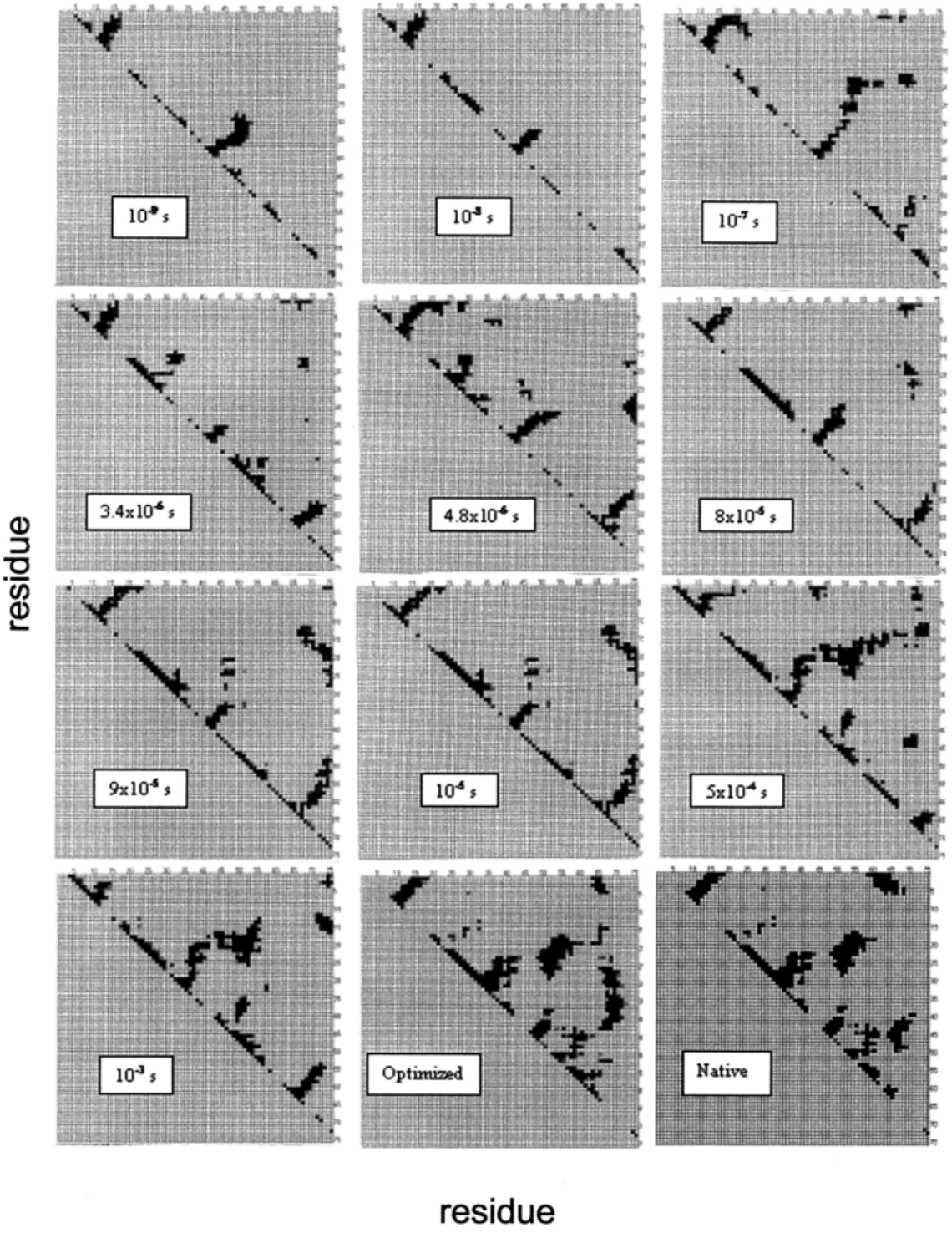

Fig. 4. Contact matrices for the folding trajectory shown in Figure 3 at the times noted. The axes are the sequence ordinals of the residues. The final, optimized contact matrix was derived by minimizing the free energy, constraining the Φ,Ψ values to remain in their current basins. The CM presented required about 100,000 steps and was halted if no alteration in the CM was detected for 10,000 steps.

the rate-limiting transition state is formed at t*. From this nucleation event, folding proceeds rapidly all the way to the native state. This dynamic behavior is seen in the contacts maps as well (Fig. 4). The early volatility through the entire polypeptide indicates that a specific intermediate does not form prior to the nucleation event.

Additionally, some bulk properties are monitored for multiple folding trajectories [Fig. 5(A,B)] including the

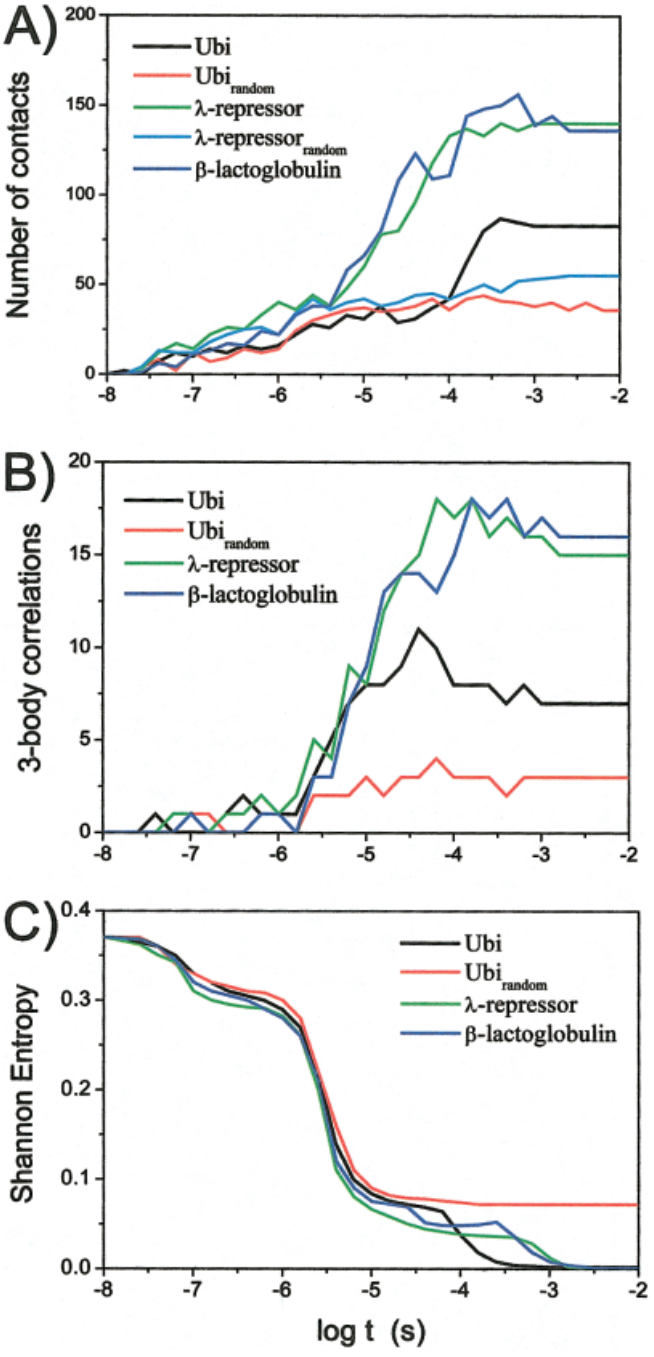

Fig. 5. Evolution of folding averaged over multiple folding trajectories. **A:** Total number of side-chain contacts formed, N(t), (α-carbon distances < 7 Å) vs. time. **B:** Number of three body correlations, $B_3(t)$, averaged over multiple runs. **C:** Normalized Shannon entropy, σ/N. The curve for the randomized version of λ-repressor is nearly identical to that for the randomized version of Ub and is not shown for clarity. The traces reflect the average of between 22 and 106 successful trajectories, depending upon the protein. The simulated folding conditions were pH 5.5, 308 K, and 10^8 iterations were performed.

average number of non-neighbor contacts, N(t), and the number of three-body correlations, $B_3(t)$, where a hydrophobic group desolvates and stabilizes a backbone-backbone hydrogen bond (see Materials and Methods). This three-body correlation reflects the environmental dependence of pairwise interactions[27–32]; the introduction of context-sensitivity of hydrogen bonds results in more expedient folding behavior.[16]

The numbers of non-local contacts and of protecting 3-body contacts have sharp transitions, but the former occurs about 1.5 decades later, at 10^{-4} sec, whereupon the native-state is formed (see below). The level of three-body correlations abruptly increases from less than unity to about eight at $10^{-5.5}$ sec. These average properties indicate that folding occurs in distinct phases at multiple time scales ranging from nsec through msec.

Shannon Entropy

The general folding pattern can be seen in the Shannon entropy, σ, a quantity that is a useful indicator of the conformational diversity of the refolding polypeptide [Figs. 5(C), 6]. Inherently, s is an average over numerous runs, and measures the diversity of backbone conformations that the polypeptide samples at a given time t resolved at the coarse LTM level of description: $\sigma(t) = -\Sigma P_i(t) \ln P_i(t)$, where $P_i(t)$ is the probability of finding the chain in the i^{th} LTM pattern. Large Shannon entropies indicate that a large fraction of conformational space is still being explored.

For a freely jointed chain with no constraints on Φ,Ψ angles, the maximum value of the Shannon entropy is the number of residues.[35] For a polypeptide, however, the torsional angles are limited by steric clash with the β-carbon. This constraint reduces the maximum value of σ by a factor of about 0.37, the ratio of the readily accessible Φ,Ψ basin areas (at T = 308 K) divided by the total area of the Φ,Ψ space, $(2\pi)^2$. After the protein has folded to the native state, the conformational diversity has been eliminated and σ is essentially zero.

About 20% of the Shannon entropy is lost in the earliest phase at 10^{-7} sec [Fig. 5(C)]. This decrease coincides with an increase in the number of contacts and three-body correlations to about 15% of their native values. In the next major phase at $10^{-5.5}$ s, fully 60% of the Shannon entropy is lost. This loss of conformational diversity drifts up to 40% of the final level upon the formation of native and non-native contacts. Even then, these contacts are itinerant, and do not correspond to formation of a persistent subset of the native contacts. Were such a subset to form, it would be the signature of a distinct, native-like intermediate. In spite of the level of contacts being only 40%, roughly the native number of three-body associations is formed in this phase. In the final folding phase, the last 20% of the Shannon entropy is lost and the final, native set of contacts is formed.

Folding of a Scrambled Sequence

To understand the sequence dependence of each of the folding phases, the algorithm was applied to 16 randomly scrambled versions of the ubiquitin sequence. The results of the folding patterns of these were unpatterned and developed no persistent structural motifs. The result from one example is presented here; the others are essentially

indistinguishable from this. Up to the final nucleation event, the same behavior is observed for this (Figs. 3 and 5) and the other scrambled versions, and for the wild-type protein. In particular, the quenching event at $10^{-5.5}$ sec is reproduced in all of them. The level of contacts and Shannon entropy of the scrambled systems at this point are nearly identical to those of the wild-type sequence, although the level of three-body correlations is lower in the random versions (see below). The scrambled sequences continue to explore the large ensemble of conformations *ad infinitum* [Fig. 5(C)] whereas the native sequence folds to the native state.

Interpretation of Early Phases

The early folding stages of Ub and its randomized counterpart exhibit a great deal of fluctuation and no persistent structure [Fig. 3(B)]. In the sub-msec interval, ubiquitin forms fluctuating secondary structures, and occasional long-range contacts. The dynamic behaviors of Ub and its randomized version are indistinguishable up to about 10 microsec (Fig. 5). These results indicate that the folding events prior to the wild-type sequence's nucleation largely reflect the adjustment of any polypeptide to a change in solvent condition.

The large decrease of the Shannon entropy for both the native and random sequence occurs with an increase in the level of three-body correlations. In the wild-type polypeptide, nucleation occurs when multiple three-body correlations form in concert, coalesce and support each other, permitting subsequent structure to condense rapidly onto the folding nucleus on the energy path down to the native structure.

Folding of Helical and Multi-State Proteins

We tested the generality of the results with a highly helical protein, λ-repressor, which also folds in a two-state manner.[21] The same phenomena observed for Ub are seen for λ-repressor and its randomized version (Fig. 5). In particular, the major reduction in the Shannon entropy occurs on the same time-scale and to the same extent for both versions of the protein. Takada and coworkers also observed similar short time behavior in a simulation of a three-helix bundle and its randomized analog.[30]

We further examined the folding process with β-lactoglobulin, a protein that folds with a distinct intermediate.[18,36] The same early phases are observed. However, two distinct quenching events are observed at $10^{-4.5}$ and $10^{-3.5}$ s. The first phase represents the formation of a specific intermediate, while the second phase represents the formation of the native structure. This result confirms that the algorithm is capable of distinguishing between the nonspecific response of the polypeptide to an environmental change, and the formation of a specific intermediate.

Differences Between Native and Scrambled Sequences

Most of the early time behavior of the native sequences also is observed in the randomized counterparts with the important exception of the number of three-body correlations. In the major burst phase at $10^{-5.5}$ s, Ub acquires the native number of correlations, or possibly even slightly more, while the random version acquires only about a third as many [Fig. 5(B)]. The difference between sequences also appears as a small difference in their respective Shannon entropies. Similar results are observed for λ-repressor (data not shown). In this sense, both the higher level of three-body correlations and the gap in the Shannon entropy indicate the formation of the native structure.

CONCLUSION

In the present simulations, no appreciable difference exists for the earliest folding events between the natural sequence proteins and their scrambled sequence counterparts. For both polypeptide classes, the number of conformations explored drops drastically, and little if any persistent structure is formed. Nevertheless, some pre-nucleation events are sensitive to the foldability of the sequence. For native proteins, the number of three-body correlations is higher while the Shannon entropy is slightly lower, following the early drastic collapse.

These results support the proposal that burst-phase, submillisecond spectroscopic signals reflect the readjustment of any polypeptide to a change in solvent condition. Foldable proteins, however, have sequences that enable three-body correlations to form both in greater number, and in a mutually reinforcing manner, which eventually results in folding to a compact, native structure. Wild-type proteins can create a self-protecting chain, while random copolymers do not. The native sequence's capability is evident even prior to the formation of persistent structure, and it can be used to distinguish foldable from non-foldable sequences.

ACKNOWLEDGMENTS

We thank Professor Walter Englander for very useful discussions. This work was supported by grants from the National Institutes of Health (T.R.S.), the National Science Foundation (R.S.B.), and Packard Foundation Interdisciplinary Science Program (T.R.S., R.S.B).

REFERENCES

1. Khorasanizadeh S, Peters ID, Roder H. Evidence for a 3-state model of protein folding from kinetic analysis of ubiquitin variants with altered core residues. Nature Struct Biol 1996;3:193–205.
2. Krantz BA, Sosnick TR. Distinguishing between two-state and three-state models for ubiquitin folding. Biochemistry 2000;39: 11696–11701.
3. Houry WA, Rothwarf DM, Scheraga HA. The nature of the initial step in the conformational folding of disulfide-intact ribonuclease a. Nature Struct Biol 1995;2:495–503.
4. Qi PX, Sosnick TR, Englander SW. The burst phase in ribonuclease a folding and solvent dependence of the unfolded state. Nature Struct Biol 1998;5:882–884.
5. Sosnick TR, Mayne L, Englander SW. Molecular collapse: The rate-limiting step in two-state cytochrome c folding. Proteins 1996;24:413–426.
6. Chan C-K, Hu Y, Takahashi S, Rousseau DL, Eaton WA, Hofrichter J. Submillisecond protein folding kinetics studied by ultrarapid mixing. Proc Natl Acad Sci USA 1997;94:1779–1784.
7. Sosnick TR, Shtilerman MD, Mayne L, Englander SW. Ultrafast

signals in protein folding and the polypeptide contracted state. Proc Natl Acad Sci USA 1997;94:8545–8550.
8. Shastry MC, Roder H. Evidence for barrier-limited protein folding kinetics on the microsecond time scale. Nature Struct Biol 1998;5: 385–392.
9. Sauder JM, Roder H. Amide protection in an early folding intermediate of cytochrome c. Fold Des 1998;3:293–301.
10. Pollack L, Tate MW, Darnton NC, Knight JB, Gruner SM, Eaton WA, Austin RH. Compactness of the denatured state of a fast-folding protein measured by submillisecond small-angle x-ray scattering. Proc Natl Acad Sci USA 1999;96:10115–10117.
11. Hagen S J, Eaton WA. Two-state expansion and collapse of a polypeptide. J Mol Biol 2000;297:781–789.
12. Akiyama S, Takahashi S, Ishimori K, Morishima I. Stepwise formation of alpha-helices during cytochrome c folding. Nature Struct Biol 2000;7:514–520.
13. Jackson SE. How do small single-domain proteins fold? Fold Des 1998;3:R81–91.
14. Fernández A, Kostov K, Berry R S. From residue matching patterns to protein folding topographies: General model and bovine pancreatic trypsin inhibitor. Proc Natl Acad Sci USA 1999;96:12991–12996.
15. Fernández A, Colubri A, Berry RS. Topology to geometry in protein folding: Beta-lactoglobulin. Proc Natl Acad Sci USA 2000;97:14062–14066.
16. Fernández A. Conformation-dependent environments in folding proteins. J Chem Phys 2001;114:2489–2502.
17. Srinivasan R, Rose GD. Linus: A hierarchic procedure to predict the fold of a protein. Proteins 1995;22:81–99.
18. Kuwata K, Shastry R, Cheng H, Hoshino M, Batt CA, Goto Y, Roder H. Structural and kinetic characterization of early folding events in beta-lactoglobulin. Nature Struct Biol 2001;8:151–155.
19. Hamada D, Segawa S, Goto Y. Non-native alpha-helical intermediate in the refolding of beta-lactoglobulin, a predominantly beta-sheet protein. Nature Struct Biol 1996;3:868–873.
20. Krantz BA, Moran LB, Kentsis A, Sosnick TR. D/H amide kinetic isotope effects reveal when hydrogen bonds form during protein folding. Nature Struct Biol 2000;7:62–71.
21. Huang GS, Oas TG. Structure and stability of monomeric lambda repressor: NMR evidence for two-state folding. Biochemistry 1995;34:3884–3892.
22. Hamada D, Goto Y. The equilibrium intermediate of beta-lactoglobulin with non-native alpha-helical structure. J Mol Biol 1997;269:479–487.
23. Fernández A, Colubri A. Microscopic dynamics from a coarsely defined solution to the protein folding problem. J Math Phys 1998;39:3167–3187.
24. Fernández A, Berry R S. Self-organization and mismatch tolerance in protein folding: General theory and an application. J Chem Phys 2000;112:5212–5222.
25. Fernández A, Colubri A, Berry RS. Topologies to geometries in protein folding: Hierarchical and nonhierarchical scenarios. J Chem Phys 2001;114:5871–5887.
26. Laskowski RA, Macarthur MW, Moss DS, Thornton JM. Procheck: A program to check the stereochemical quality of protein structures. J Appl Cryst 1993;26:283–291.
27. Minor DL Jr, Kim PS. Context is a major determinant of beta-sheet propensity. Nature 1994;371:264–267.
28. Waldburger CD, Jonsson T, Sauer RT. Barriers to protein folding: Formation of buried polar interactions is a slow step in acquisition of structure. Proc Natl Acad Sci USA 1996;93:2629–2634.
29. Liwo A, Kazmierkiewicz R, Czaplewski C, Groth M, Oldziej S, Wawak R J, Rackovsky S, Pincus MR, Scheraga HA. United-residue force field for off-lattice protein-structure simulations: Iii. Origin of backbone hydrogen-bonding cooperativity in united-residue potentials. J Comput Chem 1998;19:259–276.
30. Takada S, Luthey-Schulten Z, Wolynes P. Folding dynamics with nonadditive forces: A simulation study of a designed helical protein and a random heteropolymer. J Chem Phys 1999;110: 11616–11629.
31. Krittanai C, Johnson WC. The relative order of helical propensity of amino acids changes with solvent environment. Proteins 2000; 39:132–141.
32. Park K, Vendruscolo M, Domany E. Toward an energy function for the contact map representation of proteins. Proteins 2000;40:237–248.
33. Kentsis A, Sosnick T R. Trifluoroethanol promotes helix formation by destabilizing backbone exposure: Desolvation rather than native hydrogen bonding defines the kinetic pathway of dimeric coiled coil folding. Biochemistry 1998;37:14613–14622.
34. Fernández A, Colubri A, Berry RS. Physica A 2002;307:235–259.
35. Fernández A. The lagrangian structure of long-time torsional dynamics leading to rna folding. J Stat Phys 1998;92:237–267.
36. Forge V, Hoshino M, Kuwata K, Arai M, Kuwajima K, Batt CA, Goto Y. Is folding of beta-lactoglobulin non-hierarchic? Intermediate with native-like beta-sheet and non-native alpha-helix. J Mol Biol 2000;296:1039–1051.