nature
structural biology

Structural correspondence between the α-helix and the random-flight chain resolves how unfolded proteins can have native-like properties

Bojan Zagrovic & Vijay S Pande

Recently, we have proposed that, on average, the structure of the unfolded state of small, mostly α-helical proteins may be similar to the native structure (the 'mean-structure' hypothesis). After examining thousands of simulations of both the folded and the unfolded states of five polypeptides in atomistic detail at room temperature, we report here a result that seems at odds with the mean-structure hypothesis. Specifically, the average inter-residue distances in the collapsed unfolded structures agree well with the statistics of the ideal random-flight chain with link length of 3.8 Å (the length of one amino acid). A possible resolution of this apparent contradiction is offered by the observation that the inter-residue distances in a typical α-helix over short stretches are close to the average distances in an ideal random-flight chain.

The unfolded state of proteins is the 'other half' of the protein folding equation; however, compared to the folded state, it has received much less attention. The reasons for this are primarily the structural heterogeneity and complexity of the unfolded state and, secondarily, an understanding that biological function is predominantly carried out by the folded state. Lately, however, it has been suggested that the features of the unfolded state may have been responsible for guiding the folding process, and that this state may not be as structurally diverse as previously thought[1–7]. Indeed, Shortle and Ackerman have observed native-like topology in the denatured staphylococcal nuclease even in the presence of 8 M urea[8–10]. Furthermore, several other studies have reported a substantial presence of native-like secondary or tertiary structure in both denatured and natively unfolded molecules[5,11–15].

In a recent atomic molecular dynamics study, we have suggested that the average geometry of the unfolded state of small, mostly α-helical peptides and proteins may actually correspond to the geometry of the native state (the 'mean-structure' hypothesis)[16]. Briefly, we used distributed computing techniques[17] and a supercluster of >20,000 processors from around the world to exhaustively sample, in atomic detail and under folding conditions, the unfolded state of several small proteins from a range of structural classes: a 12-residue β-sheet tryptophan zipper[18], a 20-residue α-helical tryptophan cage[19,20], a 23-residue designed α/β BBA5 (refs. 21,22) and a 36-residue α-helical villin headpiece[23,24] (**Fig. 1**). We showed that although individual members of the unfolded ensemble in our study were substantially different from the native structure, the mean Cα-Cα and Cβ-Cβ distance matrices averaged over the entire unfolded ensemble appeared extremely native-like[16,20]. It is important to emphasize the difference between the unfolded state under folding conditions as studied in our simulations and the artificially stabilized denatured state in the presence of chemical denaturants, such as urea or guanidinium chloride, as typically studied experimentally[14].

Here we examine the mean-structure hypothesis in more detail and offer a possible explanation behind the central observation that led to its formulation. First, we recast the original observation slightly by using squared averaging of distance matrices instead of linear averaging, and show that this has negligible effect on the result. Second, we demonstrate that the unfolded-state ensembles in our study behave, on average, similarly to ideal random-flight chains with link lengths of one amino acid—a finding that is seemingly in contradiction with the central claim behind the mean-structure hypothesis. Namely, how can the average geometry of the unfolded state be native-like and random-walk-like at the same time? We resolve the apparent contradiction by showing that both the local geometry of the α-helix and the global geometry of some protein folds are similar to the average geometry of the random-flight chain with link length equivalent to the length of one amino acid. Further, we discuss our new simulations of the unfolded state of a 58-residue α-helical polypeptide protein A (**Fig. 1**). We conclude by discussing the potential implications of the close structural correspondence of the α-helix and the ideal random-flight chain.

RESULTS

Characterization of the unfolded ensembles

To avoid any native-state bias, we started our simulations from fully extended conformations (see Methods). The molecules quickly (in 5–20 ns) collapsed to the radius of gyration and the solvent-accessible surface area of the respective native states (**Table 1**)[16,20,22]. **Figure 2** captures the central finding behind the recently proposed mean-

Biophysics Program and Department of Chemistry, Stanford University, Stanford, California, 94305-5080, USA. Correspondence should be addressed to V.S.P (pande@stanford.edu).

Published online 12 October 2003; doi:10.1038/nsb995

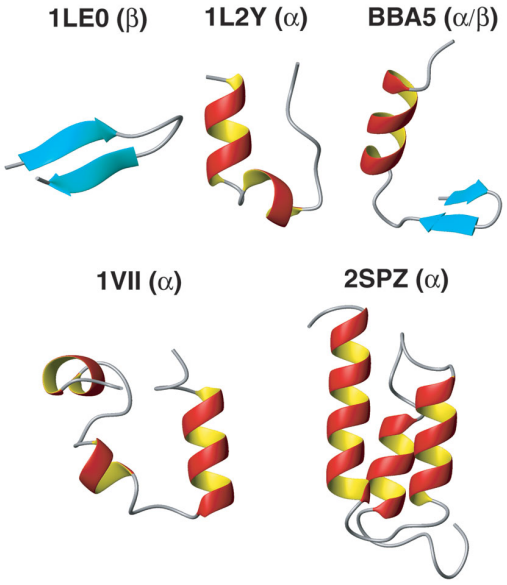


Figure 1 Three-dimensional structures as determined by NMR of the proteins used here. The secondary structure content is shown in color (α-helix, red and yellow; β-sheet, blue). The PDB entry is indicated for all proteins except BBA5, which is not in the PDB and is referred to by name throughout the manuscript.

structure hypothesis[16,20]. Briefly, we compare the unfolded ensembles tens of nanoseconds after the initiation of folding with the respective native structures in two ways. First, we calculate the distance root-mean square (dRMS) from the native Cα-Cα distance matrix for each individual member of the unfolded ensemble, and this calculation results in the depicted distributions (**Fig. 2**). In addition to the sizable average dRMS from the respective native structures (**Fig. 2**), the unfolded-state ensembles are also characterized by low native secondary structure content (**Table 1**).

Second, in analogy to typical distance-based structural techniques, such as NMR NOEs, for a given protein ensemble, we ensemble-average all of the individual Cα-Cα distance matrices, thereby generating a single unfolded-state mean-distance matrix for this ensemble. Notably, this matrix is very similar to the native distance matrix (**Fig. 2**, black arrows). In other words, even though the unfolded-state ensembles in our simulations are substantially non-native on the level of individual molecules, their overall structure is on average close to the native geometry. As explained previously[16], this holds as soon as the molecules collapse to the radius of gyration of the native state and, in general, seems to be more characteristic of α-helical than β-sheet proteins. Unlike in the original report[16], here we use r.m.s. averaging of unfolded-state distance matrices (see Methods). However, the finding also holds for linear, inverse-squared and inverse-sixth-power averaging.

To further characterize the unfolded-state ensembles in our simulations, we ask how the individual average inter-residue distances differ from those expected for a typical random-flight chain. By plotting the average Cα-Cα distances from our unfolded-state ensembles against the predictions based on the statistics of ideal random-flight chains with link length of 3.8 Å (the length of 1 amino acid)[25]:

$$\langle d^2_{i,j} \rangle^{1/2} = n^{1/2} \times 3.8 \text{ Å}$$

where n is the number of links between residues i and j (**Fig. 3**), we obtained a somewhat unexpected answer. The average inter-residue distances in our unfolded-state ensembles conform to the ideal random-flight chain behavior with a notable degree of accuracy, and this is true over a wide range of distances. Furthermore, the persistence length of this random-flight chain is exactly one amino acid, indicating extreme flexibility.

α-Helix versus random-flight chain

The above finding presents a paradox: how can the overall geometry of the peptides and proteins be both native-like and random-walk-like at the same time? Which properties of the native-state secondary and tertiary structure of these small polypeptides lead to such seemingly contradictory observations? A plausible answer and the central result of this paper is that the inter-residue distances for a typical α-helix are, over short stretches, markedly close to the corresponding average distances for an ideal random-flight chain with link length equivalent to the length of one amino acid (**Fig. 4**). In particular, the dRMS between the Cα-Cα distance matrix for a typical eight-residue α-helix and that for an average ideal random-flight chain with link length of one amino acid is only 0.8 Å (**Fig. 4a**). For comparison, the same calculation in the case of a typical eight-residue antiparallel β-sheet segment gives a dRMS of 5.4 Å. If one repeats the same analysis for every possible eight-residue peptide with repeating dihedral angles over all possible ϕ and ψ values (**Fig. 4b**), it is seen that the typical α-helix is the most similar regular structure to an ideal random-flight chain that is sterically allowed (boxed Ramachandran areas, **Fig. 4b**). In other words, the α-helix is the protein motif that is most similar to the inter-residue distances of an ideal random-flight chain.

Table 1 Details of the simulation setup and basic characterization of the unfolded ensembles

	Tryptophan zipper	Tryptophan cage	BBA5	Villin headpiece	Protein A
PDB entry	1LE0	1L2Y	BBA5	1VII	2SPZ
Number of amino acids	12	20	23	36	58
T (K)[a]	300	300	278	300	300
Time point analyzed (ns)	15	30	12	33	30
Number of structures[b]	733	1,335	4,897	4,000	1,518
$\langle R_{gyr} \rangle$ (Å)[c]	6.6 ± 0.4 (6.6)	7.4 ± 0.3 (7.5)	8.4 ± 0.7 (9.2)	10.0 ± 1.0 (9.6)	12.7 ± 1.8 (11.1)
<SASA> (Å^2)[d]	1,477 ± 72 (1,454)	1,839 ± 79 (1,795)	2,256 ± 157 (2,422)	3,085 ± 174 (3,076)	4,483 ± 326 (3,777)
<frac.SS>[e]	0.27	0.37	0.30	0.27	0.24
<frac.hel>[f]	NA	0.10	0.20	0.13	0.13

[a]Simulation temperature (T). [b]The total number of independent structures at the time point analyzed. [c]For the time point analyzed, the average radius of gyration ($\langle R_{gyr} \rangle$) with standard deviation (and the native value in parentheses). [d]For this time point, the average solvent-accessible surface area (<SASA>) with standard deviation (and the native value in parentheses). [e]The average fraction of the native-like secondary structure content (see Methods) [f]The average fraction of the native-like α-helical content (<frac.hel>, for all molecules except for the β-sheet 1LE0), as determined by DSSP[55].

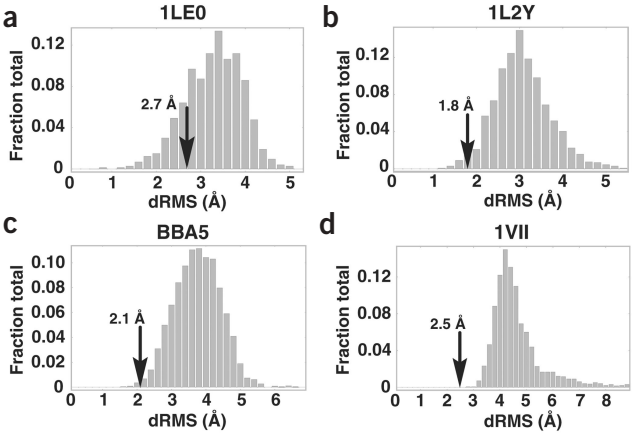


Figure 2 Distributions of dRMS from the native structures of the unfolded ensembles identified by PDB entry. (**a**) 1LE0. (**b**) 1L2Y. (**c**) BBA5. (**d**) 1VII. The means and the standard deviations of the distributions are 3.2 ± 0.6 Å (1LE0), 3.0 ± 0.6 Å (1L2Y), 3.7 ± 0.7 Å (BBA5) and 4.6 ± 1.0 Å (1VII). The arrows mark the dRMS of the mean unfolded Cα-Cα distance matrices, averaged over entire unfolded-state ensembles, from the respective native distance matrices. The fraction of individual unfolded structures that are more distant in the dRMS sense from the native structure than the mean unfolded state matrix is 78% (1LE0), 97% (1L2Y), 99% (BBA5) and 100% (1VII).

Simulations of protein A

In addition to the four small proteins and peptides discussed above, we also simulated the unfolded state of protein A, a 58-residue three-helix bundle protein, to test if the above finding holds true in this case. In the unfolded state, protein A is a highly compact globule with little native secondary structure, and its average intramolecular geometry is also described well by the statistics of the ideal random-flight chain with link length of 3.8 Å (**Fig. 5a** and **Table 1**). As seen in the case of the four smaller polypeptides, the unfolded-state mean-distance matrix is significantly closer to the native distance matrix than an average individual unfolded structure (5.8 Å versus 7.4 Å; **Fig. 5b**). However, the absolute dRMS between the mean unfolded distance matrix and the native matrix is substantial (5.8 Å), suggesting that, at least in this analysis, the mean unfolded structure does not exhibit native-like geometry.

Given the high helical content of the native protein A and taking into account the above arguments about close similarity between helices and average random-flight chains, one would still expect the unfolded state of protein A to match the native geometry in a local sense. Indeed, this is the case: if one looks at the unfolded state with a sliding window of eight residues, any such segment is on average extremely close to its native geometry (dRMS < 1 Å in most cases; **Fig. 5c**, thick black line).

Figure 3 A comparison between the individual ensemble-averaged Cα-Cα distances for the unfolded ensembles and the predictions for the ideal random-flight chain (rfc) model (by PDB entry). (**a**) 1LE0. (**b**) 1L2Y. (**c**) BBA5. (**d**) 1VII. For residues i and j in the molecule, on the x-axis is plotted the $\langle d^2_{i,j}\rangle^{1/2}_{unf}$ averaged over the entire unfolded ensemble, and on the y-axis is plotted $\langle d^2_{i,j}\rangle^{1/2}_{rfc} = n^{1/2} \times 3.8$ Å, where $n = |i - j|$ is the number of chain links between residues i and j, and 3.8 Å is the link length taken to be the typical length of one amino acid. R^2 is the Pearson correlation coefficient. The gray lines are identity lines (not fits), $y = x$, and they correspond to a situation in which the Flory's characteristic ratio, $\langle d^2_{i,j}\rangle / (nl^2)$, is exactly equal to 1. The best-fit values of this ratio in the four cases shown are only marginally different from 1 (1.06 for 1LE0, 0.98 for 1L2Y, 1.04 for BBA5 and 1.02 for 1VII).

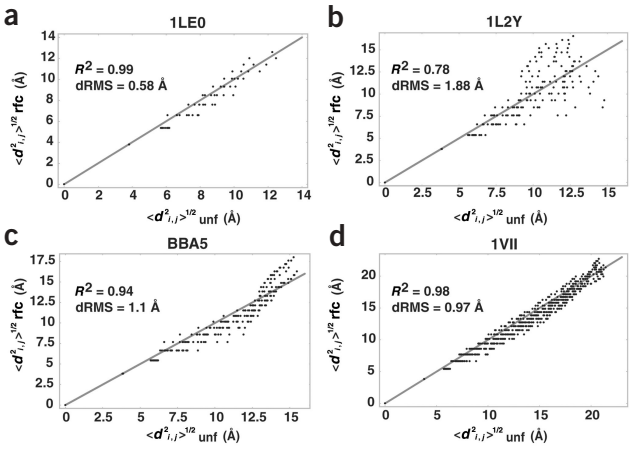
As expected, this is particularly true in the helical regions of the protein and less so in the turn regions. Furthermore, averaging of the distance matrices is again necessary for this to work—short segments belonging to individual molecules are on average significantly more different from their native geometry as compared with the corresponding ensemble-averaged segments (**Fig. 5c**, dashed gray line).

DISCUSSION

Comparison to experiment

By running thousands of simulations started from the fully extended state for a short time relative to the experimentally measured folding times, we have, in effect, captured the unfolded state under folding conditions of five small proteins and peptides. By definition, the unfolded state is an unstable, fleeting species that is difficult to study experimentally. In particular, little is known about the structure of the unfolded state under nondenaturing conditions. An NMR NOE study of the N-terminal SH3 domain of drk under nondenaturing conditions has demonstrated a compact unfolded state with both native and non-native NOEs[2]. However, an attempt at exact structural characterization yielded a range of different structures. Wright, Dyson and coworkers have examined the unfolded nondenatured state of apo plastocyanin and have also noted both native and non-native features[3,15]. Shortle and coworkers have shown that the natively disordered fragment of staphylococcal nuclease retains the topology of the native state[26–28]. In general, more is known about chemically or thermally denatured proteins. A mixture of both native (in particular α-helical) content and non-native content, with some long-range signals, is typically seen in NMR experiments[5,11–14]. However, Shortle and coworkers have studied a chemically denatured fragment of staphylococcal nuclease using dipolar couplings and provided evidence that the topology of the molecule remains highly native-like even in 8 M urea[8,10].

In our simulations, we have shown that the unfolded state under folding conditions can be native-like on average even if no individual members of the unfolded ensemble are native-like at any one time. In this study we have shown that this can be true even when the average inter-residue distances in the molecule conform to the statistics of ideal random-flight chains. It may be surprising that the unfolded collapsed globule can be well described by an idealized random-flight chain with no excluded volume. However, in the theory of globules, analogous to that of polymer melts, the excluded volume effect vanishes in collapsed globules[25]. What may be unexpected is that this effect is seen for such short polypeptides.



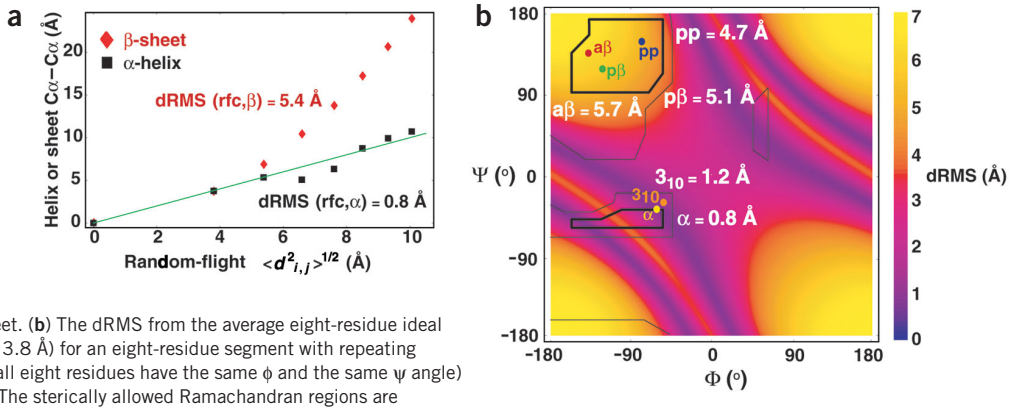


Figure 4 Helix versus random-flight chain. (**a**) Comparison of the r.m.s. inter-residue distances for an eight-residue ideal random-flight chain with link length of 3.8 Å and the Cα-Cα distances for a typical eight-residue α-helix ($\phi = -62°$, $\psi = -41°$; in black) and antiparallel β-sheet ($\phi = -139°$, $\psi = 135°$; in red). The identity line ($y = x$) is green (not a fit). The dRMS from the random-flight chain values is black for α-helix and red for antiparallel β-sheet. (**b**) The dRMS from the average eight-residue ideal random-flight chain (link length = 3.8 Å) for an eight-residue segment with repeating values of ϕ and ψ dihedral angle (all eight residues have the same ϕ and the same ψ angle) for all possible values of ϕ and ψ. The sterically allowed Ramachandran regions are outlined in black (thin line, permissive; thick line, strict). The locations of the five typical secondary structure elements and their dRMS from the average random-flight chain are outlined (α, α-helix; 3_{10}, 3_{10}-helix; aβ, antiparallel β-sheet; pβ, parallel β-sheet; pp, polyproline II helix).

In addition to the averages of the inter-residue distances presented in this study, we have also examined their distributions (B.Z. and V.S.P., unpublished data). Briefly, for separations of roughly eight residues and more, the distributions of distances adopt the expected Gaussian shape, with the expected mean and variance. For shorter separations, the distributions tend to be approximately bimodal, with the mean and the variance falling close to the idealized random-flight chain values (**Figs. 2** and **5**). Therefore, for these shorter separations (eight or fewer residues), one cannot say that the chain behaves completely in a random-flight manner but only that the average inter-residue distances of the chain can be described by the statistics of random-flight chains. Here we are concerned only with these average properties because they are the ones reported in typical ensemble-averaged structural experiments.

The question of stiffness of the chemically denatured polypeptide chain has for the past four decades been dominated by Flory and Tanford's measurements, which estimated the persistence length of a denatured protein to be 5–10 residues[29–31]. These values have recently been challenged by precise small angle X-ray scattering (SAXS) measurements that gave a significantly smaller value (~1–2 residues; K. Plaxco, personal communication)[32]. However, in our simulations, the effective persistence length of collapsed unfolded states is even lower and, in fact, extremely close to one residue. This discrepancy can be explained by the distinction between chemically denatured states, as studied experimentally, and the unfolded states under folding conditions, as modeled in our simulations. Although the exact mechanism of protein denaturation by chemical reagents such as urea or guanidinium chloride is still not fully understood, it is reasonable to expect that, by interacting with the protein backbone, these chemicals must by necessity increase the volume of the unfolded molecule, thereby increasing its effective persistence length[33]. In contrast, our simulations capture the unfolded state under folding conditions; what we see are essentially collapsed globules of the same size as the corresponding native states, with a persistence length of one amino acid.

The unfolded ensembles in our simulations collapse extremely quickly and are, on average, as compact as the corresponding native states (**Table 1**). How does this compare with experiment? Unfolded proteins under folding conditions are difficult to study experimentally, and there is little available information on their size and geometry. Flanagan *et al.* have used SAXS to study the natively disordered truncation mutant of staphylococcal nuclease and observed a highly compact unfolded state (radius of gyration only 30% larger than the native state)[34]. Shortle and coworkers have later used various spectroscopic methods to argue that the truncated staphylococcal nuclease exhibits both the size and the global topology of the native structure[8,9,26,27]. Choy *et al.*[35] have used SAXS measurements to study the unfolded state of the drk SH3 domain under native conditions. They also saw a compact unfolded state (radius of gyration 40% larger than the native state). In their explicit solvent simulation of folding of the villin headpiece molecule, Duan and

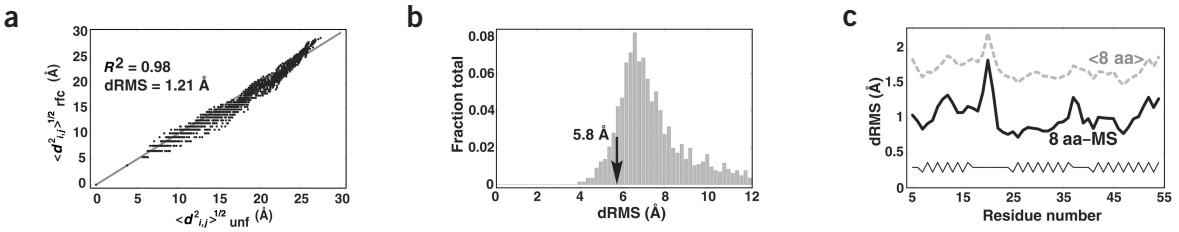


Figure 5 Analysis of the unfolded-state ensemble of protein A. (**a**) Comparison between the ensemble-averaged Cα-Cα distances for the unfolded ensemble and the predictions for the ideal random-flight chain model with link length of one amino acid (see **Fig. 2** legend for details). (**b**) Distribution of dRMS from the native structure for the unfolded state ensemble of protein A (<dRMS> = 7.4 ± 1.9 Å) and the dRMS of the mean unfolded state distance matrix from the native matrix (black arrow). This mean unfolded distance matrix is closer, in the dRMS sense, to the native structure than 87% of individual unfolded structures. (**c**) Comparison of the local structure of protein A in the native and the unfolded state (see Methods) over eight-residue stretches. We give the dRMS of the mean unfolded matrix (black thick trace) from the corresponding native matrix and the average dRMS calculated over the entire unfolded state ensemble (dashed line), for eight-residue segments centered at residues given on the *x*-axis. The secondary structure of protein A is shown as a thin black line (zigzag lines denote helical stretches).

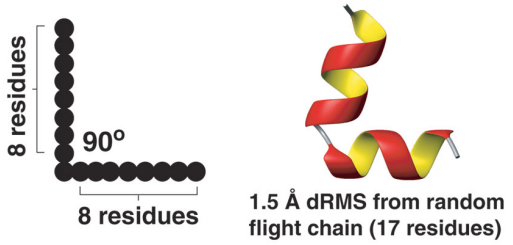


Figure 6 A structure consisting of two eight-residue α-helices, connected by a one-amino acid-linker and packed at 90° with respect to each other (the angle Cα1-Cα9-Cα17 = 90°), is only 1.5 Å-dRMS away from a 17-residue random-flight chain with a link length of 3.8 Å. This structure corresponds to an average renormalized random-flight chain with two links.

Kollman[36] have also witnessed a highly compact unfolded state with radius of gyration comparable to that of the native state. This discounts the possibility that the unfolded ensembles in our simulations are overly compact because of the implicit solvent model that we have used. Taken together, these examples suggest that the highly compact ensembles in our simulations realistically represent the unfolded state under folding conditions. Finally, the rapid collapse on the time scale of tens of nanoseconds seen in our simulations has recently received strong experimental support by the precise fluorescence measurements by Muñoz and coworkers (V. Muñoz, personal communication).

We have shown that eight consecutive residues in an α-helix are close, in terms of inter-residue distances, to an average eight-residue random-flight chain. This provides an explanation for the observation of Plaxco, Doniach and coworkers (reviewed in ref. 32) that the local structure of the polypeptide chain denatured by methanol can be highly helical, as determined by circular dichroism (CD), even though the radius of gyration of the chain exhibits random-walk-like behavior. In other words, our finding explains the local agreement between the unfolded and the native ensembles. However, the low dRMS between the average unfolded states and the native structures in the case of the tryptophan cage, BBA5 and villin mini-proteins in our simulations suggests that these unfolded molecules are, on average, native-like even when it comes to their tertiary folds. How can this be explained?

Native-state topologies can be like a random walk

Random-flight chains are characterized by the fact that their basic features, such as radius of gyration and average end-to-end distance, are not changed upon renormalization. We can treat a group of units in the chain as one individual unit in a renormalized chain with larger effective persistence length, and nothing changes. The renormalization property of random-flight chains, together with the above findings about helix geometry, suggests a hypothesis: if we place an eight-residue α-helix at right angles with another eight-residue α-helix connected by a one-residue linker, the inter-residue distances of this structure should be fairly close to the average distances in an 8 + 8 + 1 = 17-residue random-flight chain. In addition to being random-like locally, this structure should, by renormalization, globally also match the average random-flight chain with two renormalized links (90° is the expected average angle between two consecutive links of a random-flight chain). Indeed, such a helix-turn-helix motif is only 1.5-Å dRMS away from the average inter-residue distances in an ideal random-flight chain with 17 residues and link length of 3.8 Å (**Fig. 6**).

If one now looks at the structure of the tryptophan cage and BBA5 molecules, it becomes obvious that their topology is dominated by a short α-helix sandwiched at ~90° against another structural element (**Fig. 1**). In addition to their local geometry, the tertiary structure of these molecules is also already close to the average global geometry of a random-flight chain of similar length. The helix-turn-helix motif, where the two helices are packed approximately perpendicularly to each other, is commonly found in DNA-binding proteins[37], as well as in several natively unfolded proteins that exhibit binding-induced folding[38,39]. One may speculate that the binding of these proteins to their targets and/or their binding-induced folding may be facilitated by their geometry being so close to the random-flight chain on average.

The above line of reasoning can also be applied to a structure with three consecutive eight-residue helices all at right angles with each other. Namely, such a motif is expected to match the geometry of a renormalized average random-flight chain with three links. Indeed, such a structure is only 2.1-Å dRMS away from the inter-residue distances expected for an 8 + 1 + 8 + 1 + 8 = 26-residue random-flight chain. This can be applied to the case of the villin molecule because its geometry can be described approximately as three helices packed at right angles with each other.

Relationship to folding rate and stability

Recent experimental estimates of the rate of helix formation (~100–200 ns) put it close to the absolute diffusion limit[40]. Such a high rate has traditionally been explained by the fact that helix formation involves formation of sequence-local contacts only[40]. The above observation that short α-helical stretches are geometrically close to the average ideal random-flight chains suggests a new qualitative explanation for this speed. Namely, the helices form fast because the protein backbone in the unfolded, collapsed state is on average already conformationally close to the α-helical shape.

This is probably most obvious in the case of any three consecutive residues in a typical α-helix: the angle between the Cα atoms of any three consecutive residues in the α-helix is 90° (to within <1%). This is exactly the expected average angle between three consecutive nodes of a random-flight chain (triangle with sides of lengths 1, 1 and $2^{1/2}$ in reduced units). The polymeric nature of the collapsed protein backbone, just by itself, already places the molecule in the near proximity of the final α-helical shape. The final transition to the fully formed helix may then be completed by other interactions, such as the formation of hydrogen bonds and favorable van der Waals contacts. These, together with steric restraints, are probably critical for differentiating between the right-handed α-helix and other structures whose geometry is close to the random walk, such as the left-handed α-helix or the ($\phi = -150°$, $\psi = -50°$) helix (**Fig. 4b**).

The structural similarity between the average random-flight chain and the α-helix further suggests an explanation for the ubiquitous presence and stability of α-helices in different protein folds and under different environmental conditions[41]. The helix is there because any kind of thermal fluctuation that might compromise other structural motifs—that is, make random-flight chains out of them—only reinforces the helical shape on average. Further, this finding helps to rationalize the presence of early folding intermediates with non-native helical structure in some predominately β-sheet proteins[42,43]. Finally, this finding underscores the consequence of conformational averaging in defining what we know about protein structure in general. Because most, if not all, structural experiments involve time and ensemble-averaged signals, under certain circumstances, one may 'see' a helix-like structure even though it is actually a highly averaged random-flight chain.

METHODS

Simulations of the unfolded ensembles under folding conditions. Using a supercluster of >20,000 processors distributed around the world, we simulated thousands of independent trajectories, each tens of nanoseconds long, starting from a fully extended state ($\phi = -135°$, $\psi = 135°$) of five small peptides and proteins: tryptophan zipper[18], tryptophan cage[19], BBA5 (ref. 21), villin headpiece[23] and protein A[44] (**Fig. 1**). Because of the stochastic elements in the simulation algorithm (Allen's stochastic dynamics, a generalization of Langevin dynamics)[45,46], the trajectories quickly diverge from each other and independently sample the phase space.

The tryptophan zipper (PDB entry 1LE0) is a 12-residue β-hairpin peptide with a core of four tryptophans; the tryptophan cage (PDB entry 1L2Y) is a 20-residue mini-protein containing a short α-helix, a 3_{10}-helix and a C-terminal poly-proline II helix to pack against the central tryptophan; BBA5 is a 23-residue designed mini-protein consisting of an α-helix and a β-hairpin packed perpendicularly to each other; villin headpiece (PDB entry 1VII) is a 36-residue three-helix bundle protein; and staphylococcal protein A (Z domain, PDB entry 2SPZ) is a 58-residue three-helix bundle protein. All molecules except for protein A were capped with N-terminal acetyl and C-terminal amino groups, but these were excluded in the analysis. The simulations were generated using Tinker biomolecular simulation package (http://dasher.wustl.edu/tinker/), the OPLSua force field[47] and Langevin dynamics in implicit generalized Born/surface area solvent[48]. All folding simulations were run at 300 K (except for BBA5, which was run at 278 K) with water-like viscosity of $\gamma = 91$ ps^{-1}. The bonds were constrained using RATTLE[49]. No cutoffs were used for electrostatics calculations.

The structures were output once every nanosecond. All of the analyses presented here were performed on the unfolded ensembles at 15, 30, 12, 33 and 20 ns after the beginning of simulation for the tryptophan zipper, tryptophan cage, BBA5, villin headpiece and protein A, respectively. We believe that, at these time points, sufficient sampling of the unfolded basin has been carried out. However, the results presented here in no way depend on the exact choice of the time point used as long as the unfolded ensemble is fully collapsed (see below)[16,20,22]. The tryptophan cage[50], BBA5 (ref. 22), villin[51,52] and protein A[53] fold on the microsecond or slower time scale. Furthermore, by analogy to other β-sheet peptides[40,54], the tryptophan zipper is also expected to fold on the microsecond time scale. Therefore, simulations on the time scale of tens of nanoseconds captured what is essentially the unfolded state under folding conditions—that is, the folding ensemble very early into folding. More details about the simulations are given elsewhere[16,20,22].

The mean unfolded state distance matrices were generated by ensemble-averaging the distance matrices of the individual unfolded molecules. To more straightforwardly compare our unfolded ensembles with random-flight chains, this averaging was done in an r.m.s. manner, rather than in a linear manner as in the original studies[16,20]. For each individual Cα-Cα distance between residues i and j in the unfolded ensemble we found its r.m.s. average, $\langle d^2_{i,j}\rangle^{1/2}$, and these distances comprised the mean unfolded state distance matrix.

In the original study in which we proposed the mean-structure hypothesis[16], we used linear averaging of the unfolded-state distance matrices, in contrast to the squared averaging employed here. Here we use r.m.s. averaging because it allows for direct comparison with the well-known statistics of random-flight chains. The similar results obtained (see above) can be rationalized by the fact that the linear average of the inter-residue distances for an ideal random-flight chain is—the same as the r.m.s. average—also proportional to the square root of n, the number of intervening residues, and l, the link length. The two approaches are equivalent, and the arguments given here can be applied equally well in the case of linear averaging, after a slightly different value of the effective persistence length (factor of $(2/\pi)^{1/2}$) is taken into account.

The structural ensembles and individual structures throughout this study are compared using the dRMS measure defined by:

$$\mathrm{dRMS} = (2\Sigma_{i>j}\Delta d^2_{i,j} / (n^2 - \mathrm{n}))^{1/2}$$

where $\Delta d^2_{i,j} = (d_{i,j(1)} - d_{i,j(2)})^2$ is the squared difference of the distance between atoms i and j in structures 1 and 2 (in this study we use Cα-Cα distances only), and n is the number of Cα atoms in each structure.

Simulations of the native ensembles under folding conditions. To validate the stability of our simulations and explore the variability of the native basin, we also carried out extensive simulations of the native states of the five molecules studies here using the same simulation methodology and conditions as for the unfolded states described above. For each of the five molecules, we ran thousands of independent simulations starting from the experimental NMR structures for intervals of tens of nanoseconds each[16,20,22]. These simulations remained stable throughout, with respect to the radii of gyration, secondary structure content, solvent-accessible surface area and dRMS of the average structure from the experimental NMR structures[16,20]. For instance, the r.m.s.-averaged distance matrices from our native simulations at 15 ns (481 structures) for PDB entry 1LE0, at 10 ns (978 structures) for 1L2Y, at 15 ns for BBA5 (1,317 structures), at 20 ns for 1VII (1,401 structures) and at 20 ns for 2SPZ (1,190 structures) are, in the dRMS sense, only 0.7 Å (1LE0), 1.1 Å (1L2Y), 2.0 Å (BBA5), 2.2 Å (1VII) and 1.3 Å (2SPZ) away from the respective experimental NMR structures (first NMR model or average model where available). Therefore, we have used the r.m.s.-averaged native Cα-Cα distance matrices at these time points when comparing our unfolded-state ensembles with the respective native states throughout this study.

Comparisons of secondary structure. The secondary structure content of each molecule was determined using DSSP[55]. The percentage of native secondary structure content in the unfolded state was determined in the following way. Each residue was assigned to one of eight secondary structure categories from DSSP: α-helix, β-sheet, isolated bridge, 3_{10}-helix, π-helix, hydrogen bonded turn, bend or random coil. The fraction of the native secondary content was calculated as the fraction of residues in the molecule that belong to the same category in the unfolded and native structure. The fraction of the native α-helical content was calculated as the fraction of the natively helical residues in the molecule that are also helical in the unfolded structure.

When comparing the standard secondary structure elements with random-flight chains (**Fig. 4**), the following definitions were used[56]: α-helix ($\phi = -62°$, $\psi = -41°$), 3_{10}-helix ($\phi = -49°$, $\psi = -26°$), antiparallel β-sheet ($\phi = -139°$, $\psi = 135°$), parallel β-sheet ($\phi = -119°$, $\psi = 113°$) and polyproline II helix ($\phi = -78°$, $\psi = 149°$).

Average distances for random-flight chains. Throughout this study we compared the properties of unfolded ensembles and of the typical structural elements in proteins with the properties of random-flight chains. The r.m.s.-averaged inter-residue distance between residues i and j in a random-flight chain is calculated according to[25]:

$$\langle d^2_{i,j}\rangle^{1/2} = n^{1/2} \times l$$

where $n = |i - j|$ is the number of chain links between residues i and j, and l is the link length of the chain in units of Å (taken to be 3.8 Å here).

Local structure analysis for protein A. For each segment of eight consecutive residues centered at a given residue (x-axis, **Fig. 5c**), we built an 8×8 Cα-Cα distance matrix and then calculated the dRMS between this matrix and the corresponding 8×8 native matrix. If we repeat this for each member of the unfolded-state ensemble of protein A, we get a distribution (mean plotted as a gray dashed line, **Fig. 5c**). However, if for each central residue we first find a mean 8×8 matrix by averaging all of the corresponding matrices for that segment over the entire unfolded-state ensemble, the result is significantly more native-like (dRMS of the mean structure given in black, **Fig. 5c**). This procedure is identical to finding the mean of the distribution and the dRMS of the mean distance matrix from the native matrix (**Figs. 2** and **5b**), but here we apply it to eight residue segments only.

Structure generation. Structures in **Figures 4b** and **6** were generated using the OPLSua geometric parameters[47].

ACKNOWLEDGMENTS

We thank all the contributors to the Folding@Home project (a complete list can be found at http://folding.stanford.edu). B.Z. acknowledges support from the Howard Hughes Medical Institute Predoctoral Fellowship program. We thank R. Baldwin, H. Andersen, M. Levitt and members of the Pande group for help and discussions. This work was supported by grants from the ACS PRF, US National Institutes of Health (NIH), US National Science Foundation MRSEC CPIMA, NIH BISTI, ARO and Stanford University (Internet 2), as well as by gifts from the Intel and Google corporations. B.Z. dedicates this work to his grandmother Katica.

COMPETING INTERESTS STATEMENT.

The authors declare that they have no competing financial interests.

Received 12 December 2002; accepted 26 August 2003
Published online at http://www.nature.com/naturestructuralbiology/

1. Plaxco, K.W. & Gross, M. Unfolded, yes, but random? Never! *Nat. Struct. Biol.* **8**, 659–660 (2001).
2. Mok, Y.K., Kay, C.M., Kay, L.E. & Forman-Kay, J. NOE data demonstrating a compact unfolded state for an SH3 domain under non-denaturing conditions. *J. Mol. Biol.* **289**, 619–638 (1999).

3. Bai, Y., Chung, J., Dyson, H.J. & Wright, P.E. Structural and dynamic characterization of an unfolded state of poplar apo-plastocyanin formed under nondenaturing conditions. *Protein Sci.* **10**, 1056–1066 (2001).
4. Fersht, A.R. & Daggett, V. Protein folding and unfolding at atomic resolution. *Cell* **108**, 573–582 (2002).
5. Klein-Seetharaman, J. *et al.* Long-range interactions within a nonnative protein. *Science* **295**, 1719–1722 (2002).
6. Tollinger, M., Skrynnikov, N.R., Mulder, F.A., Forman-Kay, J.D. & Kay, L.E. Slow dynamics in folded and unfolded states of an SH3 domain. *J. Am. Chem. Soc.* **123**, 11341–11352 (2001).
7. van Gunsteren, W.F., Burgi, R., Peter, C. & Daura, X. The key to solving the protein-folding problem lies in an accurate description of the denatured state. *Angew. Chem. Int. Ed. Engl.* **40**, 352–355 (2001).
8. Shortle, D. & Ackerman, M.S. Persistence of native-like topology in a denatured protein in 8 M urea. *Science* **293**, 487–489 (2001).
9. Shortle, D. The expanded denatured state: an ensemble of conformations trapped in a locally encoded topological space. *Adv. Protein Chem.* **62**, 1–23 (2002).
10. Ackerman, M.S. & Shortle, D. Robustness of the long-range structure in denatured staphylococcal nuclease to changes in amino acid sequence. *Biochemistry* **41**, 13791–13797 (2002).
11. Schwarzinger, S., Wright, P.E. & Dyson, H.J. Molecular hinges in protein folding: the urea-denatured state of apomyoglobin. *Biochemistry* **41**, 12681–12686 (2002).
12. Schwarzinger, S., Kroon, G.J., Foss, T.R., Wright, P.E. & Dyson, H.J. Random coil chemical shifts in acidic 8 M urea: implementation of random coil shift data in NMRView. *J. Biomol. NMR* **18**, 43–48 (2000).
13. Lietzow, M.A., Jamin, M., Dyson, H.J. & Wright, P.E. Mapping long-range contacts in a highly unfolded protein. *J. Mol. Biol.* **322**, 655–662 (2002).
14. Lecomte, J.T. & Falzone, C.J. Where U and I meet. *Nat. Struct. Biol.* **6**, 605–608 (1999).
15. Dyson, H.J. & Wright, P.E. Insights into the structure and dynamics of unfolded proteins from nuclear magnetic resonance. *Adv. Protein Chem.* **62**, 311–340 (2002).
16. Zagrovic, B., Snow, C., Khaliq, S., Shirts, M. & Pande, V. Native-like mean structure in the unfolded ensemble of small proteins. *J. Mol. Biol.* **323**, 153–164 (2002).
17. Zagrovic, B., Sorin, E.J. & Pande, V. β-hairpin folding simulations in atomistic detail using an implicit solvent model. *J. Mol. Biol.* **313**, 151–169 (2001).
18. Cochran, A.G., Skelton, N.J. & Starovasnik, M.A. Tryptophan zippers: stable, monomeric β-hairpins. *Proc. Natl. Acad. Sci. USA* **98**, 5578–5583 (2001).
19. Neidigh, J.W., Fesinmeyer, R.M. & Andersen, N.H. Designing a 20-residue protein. *Nat. Struct. Biol.* **9**, 425–430 (2002).
20. Snow, C.D., Zagrovic, B. & Pande, V.S. The trp cage: folding kinetics and unfolded state topology via molecular dynamics simulations. *J. Am. Chem. Soc.* **124**, 14548–14549 (2002).
21. Struthers, M., Ottesen, J.J. & Imperiali, B. Design and NMR analyses of compact, independently folded BBA motifs. *Fold. Des.* **3**, 95–103 (1998).
22. Snow, C.D., Nguyen, H., Pande, V.S. & Gruebele, M. Absolute comparison of simulated and experimental protein-folding dynamics. *Nature* **420**, 102–106 (2002).
23. McKnight, C.J., Matsudaira, P.T. & Kim, P.S. NMR structure of the 35-residue villin headpiece subdomain. *Nat. Struct. Biol.* **4**, 180–184 (1997).
24. Zagrovic, B., Snow, C.D., Shirts, M.R. & Pande, V.S. Simulation of folding of a small α-helical protein in atomistic detail using worldwide-distributed computing. *J. Mol. Biol.* **323**, 927–937 (2002).
25. Grosberg, A.Y. & Khokhlov, A.R. *Statistical Physics of Macromolecules* (American Institutes of Physics, New York, 1994).
26. Gillespie, J.R. & Shortle, D. Characterization of long-range structure in the denatured state of staphylococcal nuclease. I. Paramagnetic relaxation enhancement by nitroxide spin labels. *J. Mol. Biol.* **268**, 158–169 (1997).
27. Gillespie, J.R. & Shortle, D. Characterization of long-range structure in the denatured state of staphylococcal nuclease. II. Distance restraints from paramagnetic relaxation and calculation of an ensemble of structures. *J. Mol. Biol.* **268**, 170–184 (1997).
28. Zhang, O., Kay, L.E., Shortle, D. & Forman-Kay, J.D. Comprehensive NOE characterization of a partially folded large fragment of staphylococcal nuclease Δ131Δ, using NMR methods with improved resolution. *J. Mol. Biol.* **272**, 9–20 (1997).
29. Brant, D.A. & Flory, P.J. The configuration of random polypeptide chain. I. Experimental results. *J. Am. Chem. Soc.* **87**, 2788–2791 (1965).
30. Brant, D.A. & Flory, P.J. The configuration of random polypeptide chain. II. Theory. *J. Am. Chem. Soc.* **87**, 2791–2800 (1965).
31. Tanford, C. Protein denaturation. *Adv. Protein Chem.* **23**, 121–282 (1968).
32. Millett, I.S., Doniach, S. & Plaxco, K.W. Toward a taxonomy of the denatured state: small angle scattering studies of unfolded proteins. *Adv. Protein Chem.* **62**, 241–262 (2002).
33. Schellman, J.A. Fifty years of solvent denaturation. *Biophys. Chem.* **96**, 91–101 (2002).
34. Flanagan, J.M., Kataoka, M., Shortle, D. & Engelman, D.M. Truncated staphylococcal nuclease is compact but disordered. *Proc. Natl. Acad. Sci. USA* **89**, 748–752 (1992).
35. Choy, W.Y. *et al.* Distribution of molecular size within an unfolded state ensemble using small-angle X-ray scattering and pulse field gradient NMR techniques. *J. Mol. Biol.* **316**, 101–112 (2002).
36. Duan, Y. & Kollman, P.A. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **282**, 740–744 (1998).
37. Branden, C.I. & Tooze, J. *Introduction to Protein Structure.* (Garland Publishing, New York, 1999).
38. Radhakrishnan, I. *et al.* Solution structure of the KIX domain of CBP bound to the transactivation domain of CREB: a model for activator:coactivator interactions. *Cell* **91**, 741–752 (1997).
39. Demarest, S.J. *et al.* Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators. *Nature* **415**, 549–553 (2002).
40. Eaton, W.A. *et al.* Fast kinetics and mechanisms in protein folding. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 327–359 (2000).
41. Baldwin, R.L. & Rose, G.D. Is protein folding hierarchic? I. Local structure and peptide folding. *Trends Biochem. Sci.* **24**, 26–33 (1999).
42. Forge, V. *et al.* Is folding of β-lactoglobulin non-hierarchic? Intermediate with native-like β-sheet and non-native α-helix. *J. Mol. Biol.* **296**, 1039–1051 (2000).
43. Kuwata, K. *et al.* Structural and kinetic characterization of early folding events in β-lactoglobulin. *Nat. Struct. Biol.* **8**, 151–155 (2001).
44. Tashiro, M. *et al.* High-resolution solution NMR structure of the Z domain of staphylococcal protein A. *J. Mol. Biol.* **272**, 573–590 (1997).
45. Allen, M.P. Brownian dynamics simulation of a chemical reaction in solution. *Mol. Phys.* **40**, 1073–1087 (1980).
46. Allen, M.P. & Tildesley, D.J. *Computer Simulations of Liquids.* (Oxford University Press, Oxford, 1987).
47. Jorgensen, W.L. & Tirado-Rives, J. The OPLS potential functions for proteins: energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **110**, 1666–1671 (1988).
48. Qiu, D., Shenkin, P.S., Hollinger, F.P. & Still, W.C. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. *J. Phys. Chem.* 3005–3014 (1997).
49. Andersen, H.C. Rattle: a 'velocity' version of the Shake algorithm for molecular dynamics calculations. *J. Comp. Phys.* **52**, 24–34 (1983).
50. Qiu, L., Pabit, S.A., Roitberg, A.E. & Hagen, S.J. Smaller and faster: the 20-residue Trp-cage protein folds in 4 microseconds. *J. Am. Chem. Soc.* **124**, 12952–12953 (2002).
51. Wang, M. *et al.* Dynamic NMR Line-shape analysis demonstrates that the villin headpiece subdomain folds on the microsecond time scale. *J. Am. Chem. Soc.* **125**, 6032–6033 (2003).
52. Kubelka, J., Eaton, W.A. & Hofrichter, J. Experimental tests of villin subdomain folding simulations. *J. Mol. Biol.* **329**, 625–630 (2003).
53. Myers, J.K. & Oas, T.G. Preorganized secondary structure as an important determinant of fast protein folding. *Nat. Struct. Biol.* **8**, 552–558 (2001).
54. Muñoz, V., Thompson, P.A., Hofrichter, J. & Eaton, W.A. Folding dynamics and mechanism of β-hairpin formation. *Nature* **390**, 196–199 (1997).
55. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
56. Creighton, T.E. *Proteins: Structure and Molecular Properties.* (W.H. Freeman, New York, 1993).