

## Native-like Mean Structure in the Unfolded Ensemble of Small Proteins

Bojan Zagrovic<sup>1</sup>, Christopher D. Snow<sup>1</sup>, Siraj Khaliq<sup>2</sup>  
Michael R. Shirts<sup>2</sup> and Vijay S. Pande<sup>1,2\*</sup>

<sup>1</sup>*Biophysics Program  
Stanford University, Stanford  
CA 94305-5080, USA*

<sup>2</sup>*Department of Chemistry  
Stanford University, Stanford  
CA 94305-5080, USA*

The nature of the unfolded state plays a great role in our understanding of proteins. However, accurately studying the unfolded state with computer simulation is difficult, due to its complexity and the great deal of sampling required. Using a supercluster of over 10,000 processors we have performed close to 800  $\mu$ s of molecular dynamics simulation in atomistic detail of the folded and unfolded states of three polypeptides from a range of structural classes: the all-alpha villin headpiece molecule, the beta hairpin tryptophan zipper, and a designed alpha-beta zinc finger mimic. A comparison between the folded and the unfolded ensembles reveals that, even though virtually none of the individual members of the unfolded ensemble exhibits native-like features, the mean unfolded structure (averaged over the entire unfolded ensemble) has a native-like geometry. This suggests several novel implications for protein folding and structure prediction as well as new interpretations for experiments which find structure in ensemble-averaged measurements.

© 2002 Elsevier Science Ltd. All rights reserved

*Keywords:* mean-structure hypothesis; unfolded state of proteins; distributed computing; conformational averaging

\*Corresponding author

### Introduction

Historically, the unfolded state of proteins has received significantly less attention than the folded state.<sup>1</sup> The reasons for this are primarily its structural heterogeneity and complexity, and secondarily a belief that biological function is predominantly mediated by the native state. The molten globule, as a specific example of a non-folded state, has been studied somewhat more intensively, but little is known about its structure.<sup>2</sup> Several recent studies of the chemically or thermally denatured proteins, both experimental and theoretical, have suggested that the structure of the denatured state may not be as diverse as previously thought, and that long-range order in the denatured state may play a role in defining protein folding mechanism.<sup>3–13</sup> The majority of these studies have focused on the properties of the artificially generated non-native samples, and there has in general been very little focus on the structure and the dynamics of the unfolded state

under folding conditions, with some notable exceptions.<sup>14–16</sup> This is understandable since under such conditions the unfolded state is an unstable, fleeting species making any kind of quantitative experimental measurement very difficult. Here, it is important to emphasize the distinction between the unfolded state, a transient species *en route* to the folded state, and the denatured state, an artificially stabilized non-native state. For instance, in a typical stop-flow folding experiment, the denatured state refers to the protein in the presence of urea or guanidinium chloride, while the unfolded state refers to the same species after the denaturant has been diluted out and the protein is beginning to fold. The structural and dynamic differences between the two species have been noted before.<sup>15</sup>

Computer simulations of either the unfolded or the denatured state have so far been limited by the immense computational power required for accurate sampling. The unfolded state is in fact a greater challenge to simulate using conventional means than the folded state precisely because of its structural diversity. While there have been several simulations of the denatured state before, in particular high temperature denaturing studies,<sup>4,6,7,12,17–19</sup> it has been debated whether the sampling has been sufficient. Indeed, most studies

Abbreviations used: dRMS, distance-based root-mean square deviation; GB/SA, generalized Born/surface area.

E-mail address of the corresponding author:  
[pande@stanford.edu](mailto:pande@stanford.edu)

**Table 1.** Summary of the native equilibrium simulations

|   | Native villin  | Native TrpZip | Native BBA5    |
|---|----------------|---------------|----------------|
| Temperature (K)   | 300            | 278           | 278            |
| Total time ( $\mu$ s)   | 90.6           | 19.3          | 72.0           |
| Initial $R_{\text{gyr}}$ ( $\text{\AA}$ )                             | 9.6            | 6.6           | 9.2            |
| Initial SASA ( $\text{\AA}^2$ )                                       | 3076           | 1449          | 2422           |
| Representative time point (ns) (no. structures)                       | 20 (1401)      | 15 (481)      | 15 (1317)      |
| $\langle R_{\text{gyr}} \rangle$ at $t_{\text{rep}}$ ( $\text{\AA}$ ) | $9.8 \pm 0.8$  | $6.6 \pm 0.2$ | $8.6 \pm 0.7$  |
| $\langle \text{SASA} \rangle$ at $t_{\text{rep}}$ ( $\text{\AA}^2$ )  | $3027 \pm 137$ | $1454 \pm 57$ | $2256 \pm 115$ |
| SS at $t_{\text{rep}}$ (%)  | 87             | 78            | 90             |
| $C^\alpha$ -dRMS <sub>ms</sub> at $t_{\text{rep}}$ ( $\text{\AA}$ )   | 1.5            | 0.6           | 2.0            |

Representative time point ( $t_{\text{rep}}$ ) indicates the time at which we have calculated the mean native structures used for comparison with the unfolded simulations throughout this article. The number of independent structures used in these averages is shown in the  $t_{\text{rep}}$  row. The secondary structure at  $t_{\text{rep}}$  (SS at  $t_{\text{rep}}$ ) refers to the fraction of the native ensemble at  $t_{\text{rep}}$ , which has 14 or more helical residues in the case of villin, four or more beta-sheet residues in the case of TrpZip, and four or more helical residues and two or more beta-sheet residues in the case of BBA5, indicating stable secondary structure. All secondary structure content is determined using DSSP.<sup>40</sup>  $C^\alpha$ -dRMS<sub>ms</sub> at  $t_{\text{rep}}$  refers to the  $C^\alpha$ -dRMS of the mean native structure at  $t_{\text{rep}}$  from the initial experimental native structure used in the simulations.  $R_{\text{gyr}}$  refers to the radius of gyration, and SASA refers to the solvent accessible surface area as determined by DSSP. We show values for the two for both the initial structure and the ensemble at the  $t_{\text{rep}}$ .  $\langle \rangle$  brackets refer to ensemble averages throughout the text. Error values refer to the standard deviation around the mean of a given population.

employ a few (one to ten) simulations on the nanosecond timescale. In addition, it is also debatable how relevant are the results obtained under high temperature conditions for our understanding of the unfolded state under folding conditions. Recently, we have introduced a novel computational approach aimed at addressing the issue of vastly improving the sampling in protein simulations: for our calculations we have employed distributed computing techniques and a super-cluster of more than 10,000 processors.<sup>20,21</sup> Using this computational resource, we have run thousands of fully independent atomistic molecular dynamics (MD) trajectories starting from the fully extended state of three small proteins, each tens of nanoseconds long (see Methods; Figures 1(a), 2(a) and 3(a)). In addition, as an important control for our folding simulations, we have also run thousands of trajectories starting from the experimental native structures of the three molecules.

The advantage of performing a large number of independent, relatively short folding simulations is twofold. First, by the stochastic nature of the folding process and exponential kinetics, in an ensemble consisting of thousands of such trajectories we can expect to observe a small but significant number of folding events which on average would take much longer to occur.<sup>21</sup> Specifically, in an ensemble of 10,000 trajectories, each of which is 10 ns long, one expects to see about ten folding events for a protein that folds with single exponential kinetics and time constant of 10  $\mu$ s. Second, such an approach gives us a detailed picture of the unfolded ensemble very early into folding (tens of nanoseconds after initiation of folding). Here we have focused on this latter aspect of our data for three different polypeptides: a 36 residue three-helix bundle villin headpiece protein,<sup>22</sup> a 12 residue  $\beta$ -hairpin tryptophan zipper peptide (TrpZip<sup>23</sup>), and a 23 residue designed  $\beta\beta\alpha$  zinc finger mimic (BBA5<sup>24</sup>). Our aggregate simulation time is nearly

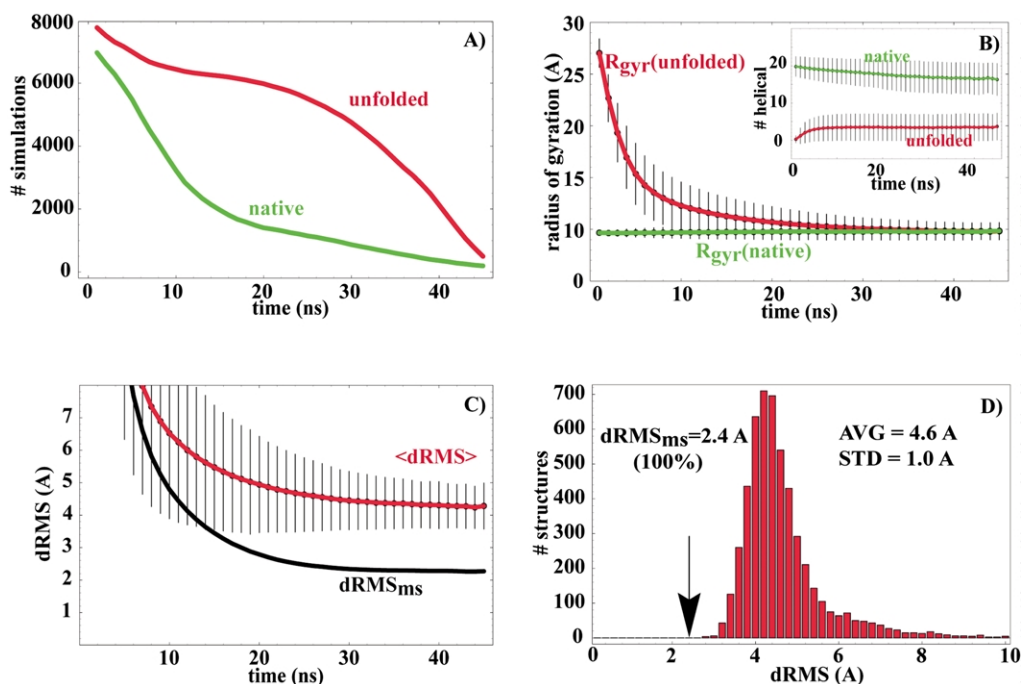
one millisecond, orders of magnitude larger than previous atomistic MD simulations.<sup>25,26</sup> This unprecedented sampling has allowed us to observe previously inaccessible effects, which we describe below. The central question that we ask is what is the structure of the unfolded state on average.

## Results

The folding simulations for all three molecules studied were started from extended conformations. In approximately 10 ns (TrpZip and BBA5) or 20 ns (villin) the unfolded ensembles non-specifically collapse to form compact conformations. On average, these conformations exhibit native-like radii of gyration and solvent accessible surface areas (Table 1, Figures 1(b), 2(b) and 3(b)). Simulations started from the folded structures, in contrast, remain stable throughout, with respect to the radii of gyration, secondary structure content, solvent accessible surface area, and distance-based root-mean square deviation of the average structure (dRMS) (Table 1, Figures 1(b), 2(b) and 3(b)) from the experimental structures. In Figures 5–7 (see below), we compare the experimental structures of the three molecules with the representative native structures from our simulations (Figures 5(a) and (b), 6(a) and (b), and 7(a) and (b)), and the similarity is obvious. This attests to the stability of our simulations and suggests that our simulated native ensembles could be used for comparison with the unfolded ensembles.

Inspection of individual members of the unfolded ensembles reveals very heterogeneous populations. Except for the majority of unfolded molecules being collapsed, they are structurally quite diverse. Despite this diversity, a meaningful question to ask is what do these heterogeneous ensembles look like on average. A natural first question is how do the averaged properties of the

## Villin



**Figure 1.** Villin simulations. (a) The total number of independent simulations that have reached a given time point for the native ensemble (green) and the unfolded ensemble (red). (b) The average radius of gyration,  $\langle R_{\text{gyr}} \rangle$  over time for the native ensemble (green) and the unfolded ensemble (red). Inset is the average  $\alpha$ -helical content per molecule as determined by DSSP for the two ensembles. (c) Comparison between the ensemble-averaged dRMS from the mean native matrix at 20 ns for the unfolded ensemble over time,  $\langle \text{dRMS} \rangle$ , (red) and the dRMS of the mean unfolded structure from the mean native matrix,  $\text{dRMS}_{\text{ms}}$ , (black). (d) Distribution of dRMS from the mean native matrix at 20 ns for all individual unfolded molecules at the 27 ns time point. The arrow marks the dRMS from the mean native matrix at 20 ns of the mean matrix on the basis of the entire unfolded ensemble at 27 ns ( $\text{dRMS}_{\text{ms}}$ , where subscript ms denotes “mean structure”). We indicate the  $\text{dRMS}_{\text{ms}}$  and the percentage of individual unfolded structures that are more different in the dRMS sense from the mean native matrix than the mean unfolded matrix. The mean (AVG) and the standard deviation (STD) of the distribution shown are also indicated. All dRMS values refer to  $C^\alpha$ -dRMS. In all Figures, the error bars represent the standard deviation around the mean at the given time point; all dRMS distributions are binned with 0.2 Å resolution.

unfolded state compare with the native state? Experimental techniques such as NMR, EPR and FRET, which provide information on interatomic distances, typically involve ensemble-averaged signals. In such experiments, each observed signal corresponding to a particular pair of nuclei, spin labels or chromophores (depending on the experiment) is independently averaged over the entire ensemble. We took a similar approach here in order to analyze the average features of the structurally diverse unfolded ensembles seen computationally. Namely, we calculated a distance matrix over all possible  $C^\alpha$  (or  $C^\beta$ ) pairs for each individual structure in the unfolded ensemble at a given time point, and then averaged these matrices to obtain one mean distance matrix, representative of the entire ensemble at that time point. We have repeated this calculation for the equilibrium simulations started from the folded states of the three polypeptides listed above. This has allowed for a natural comparison between the folded and unfolded ensembles through dRMS calculations. Namely, a standard way of comparing two dis-

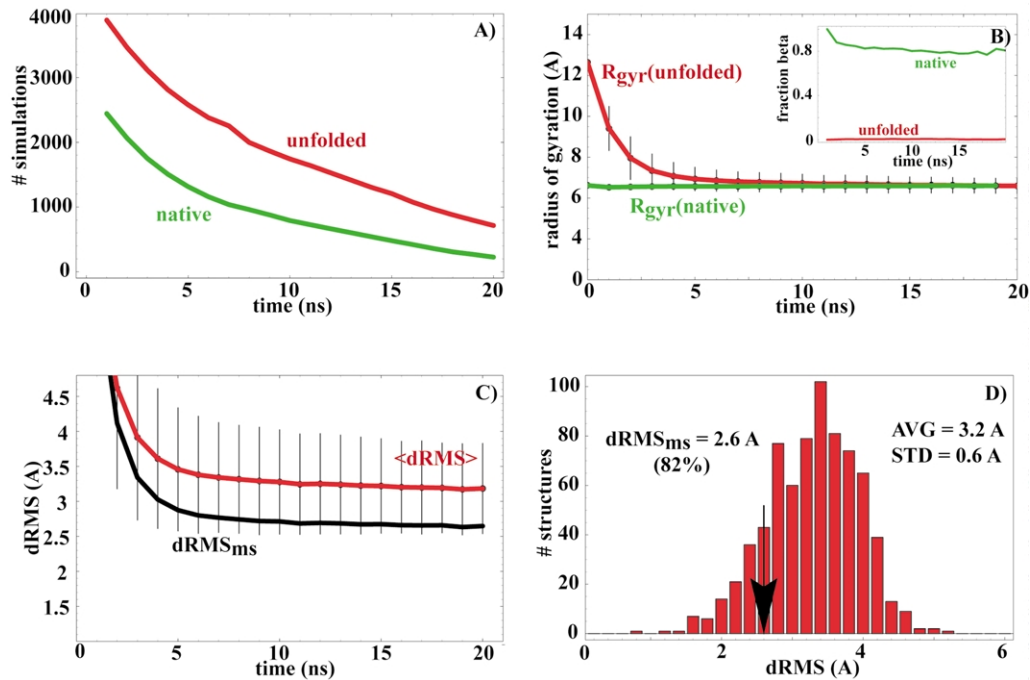
tance matrices (i.e. two structures) involves calculating distance root-mean square deviation:

$$\text{dRMS} = \sqrt{2 \frac{\sum_{i>j} [D_{ij}(1) - D_{ij}(2)]^2}{n(n-1)}}$$

where  $D_{ij}(x)$  refers to the distance between atoms  $i$  and  $j$  in structure  $x$ , and  $n$  is the total number of atoms included within each structure. The results presented here are on the basis of linear averaging of distance matrices, which we believe is statistically most natural. We have also examined the effects of  $\langle r^{-6} \rangle^{-1/6}$  averaging, intrinsic to dipolar-coupling experiments, as well as  $\langle r^{-2} \rangle^{-1/2}$  averaging, but no significant differences were seen. The conclusions presented here are qualitatively insensitive to the nature of the averaging.

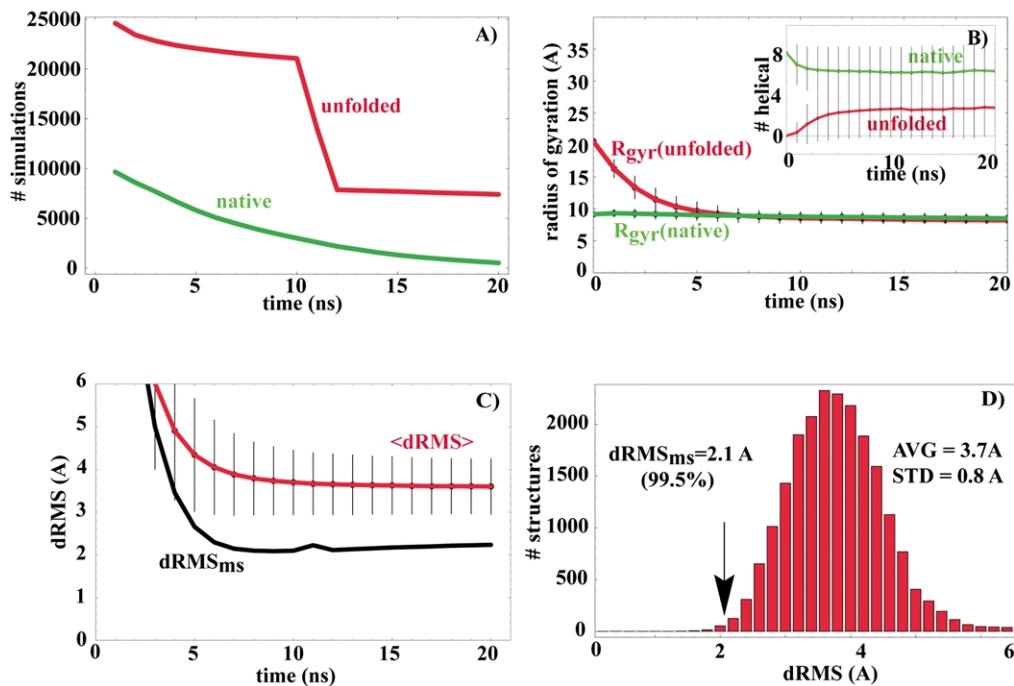
In Figures 1–3, we contrast the unfolded ensembles at different time points with their respective simulated native ensembles at the time points indicated. The top curves in Figures 1(c), 2(c) and 3(c) show the ensemble-averaged  $C^\alpha$ -dRMS

## Tryptophan Zipper



**Figure 2.** Tryptophan zipper simulations. (a)–(d) The same type of data as described for Figure 1 legend for villin (Figure 1(a)–(d), respectively), with the following modifications. Inset in (b) We show the ensemble-averaged secondary structure content of the native and the unfolded ensembles of tryptophan zipper over time: we show the fraction of the two ensembles which has four or more beta-sheet residues as determined by DSSP. In (d) we compare the unfolded ensemble at 20 ns with the mean folded matrix at 15 ns.

## BBA5



**Figure 3.** BBA5 simulations. (a)–(d) The same type of data as described for Figure 1 legend for villin (Figure 1(a)–(d), respectively), with the following modifications. Inset in (b) we show the ensemble-averaged helical secondary structure content of the native and the unfolded ensembles of BBA5 over time as determined by DSSP. In (d) we compare the unfolded ensemble at 10 ns with the mean folded structure at 15 ns.

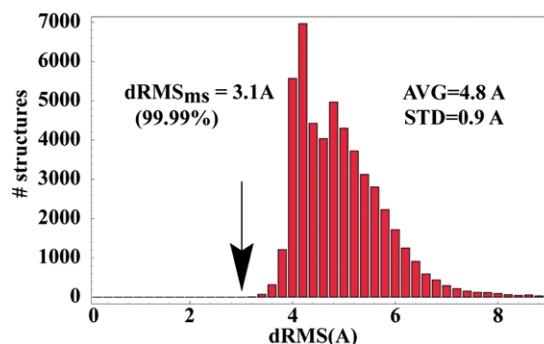
**Table 2.** Summary of the simulations started from the extended state

|   | Unfolded villin | Unfolded TrpZip | Unfolded BBA5  |
|---|-----------------|-----------------|----------------|
| Temperature (K)   | 300             | 278             | 278            |
| Total time ( $\mu$ s)   | 225.6           | 42.7            | 337.2          |
| Representative time point (ns) (no. structures)   | 27 (5248)       | 20 (715)        | 10 (21,068)    |
| $\langle \text{dRMS} \rangle_{\text{C}\alpha}$ ( $\text{\AA}$ ) versus initial native   | $5.0 \pm 1.9$   | $3.5 \pm 0.6$   | $4.3 \pm 0.8$  |
| $\text{C}^\alpha$ -dRMS <sub>ms</sub> ( $\text{\AA}$ ) versus initial native (% better) | 3.2 (99.7)      | 3.0 (55)        | 3.0 (95.2)     |
| $\langle R_{\text{gyr}} \rangle$ at $t_{\text{rep}}$ ( $\text{\AA}$ )                   | $10.2 \pm 1.2$  | $6.6 \pm 0.4$   | $8.6 \pm 0.8$  |
| $\langle \text{SASA} \rangle$ at $t_{\text{rep}}$ ( $\text{\AA}^2$ )                    | $3122 \pm 193$  | $1477 \pm 72$   | $2256 \pm 157$ |
| SS at $t_{\text{rep}}$ (%)  | 0.6             | 0.7             | 5              |
| $\langle \text{dRMS} \rangle_{\text{C}\beta}$ ( $\text{\AA}$ )                          | $5.1 \pm 1.9$   | $3.4 \pm 0.6$   | $4.1 \pm 0.7$  |
| $\text{C}^\beta$ -dRMS <sub>ms</sub> ( $\text{\AA}$ ) (% better)                        | 2.7 (100)       | 2.6 (92)        | 2.3 (99.8)     |

Representative time point ( $t_{\text{rep}}$ ) refers to the time point in folding simulations for which we have shown complete dRMS distributions in Figures 1(d), 2(d) and 3(d). The number of structures in the  $t_{\text{rep}}$  row refers to the total number of independent structures at  $t_{\text{rep}}$ . The values at  $t_{\text{rep}}$  are representative of the values at other time points after the collapse. We show ensemble-averaged  $\text{C}^\alpha$ -dRMS of the unfolded ensemble at  $t_{\text{rep}}$  from the initial structure used in native state simulations ( $\langle \text{dRMS} \rangle_{\text{C}\alpha}$  versus initial native), and the  $\text{C}^\alpha$ -dRMS of the mean unfolded structure at  $t_{\text{rep}}$  from the initial structure ( $\text{C}^\alpha$ -dRMS<sub>ms</sub> versus initial native). We also show ensemble-averaged  $\text{C}^\beta$ -dRMS of the unfolded ensemble at  $t_{\text{rep}}$  from the mean native structure at times indicated in Table 1 ( $\langle \text{dRMS} \rangle_{\text{C}\beta}$ ), and the  $\text{C}^\beta$ -dRMS of the mean unfolded structure at  $t_{\text{rep}}$  from the mean native structure ( $\text{C}^\beta$ -dRMS<sub>ms</sub>). % better refers to the percentage of unfolded state ensemble at  $t_{\text{rep}}$ , which is more distant from the native structure in the dRMS sense than the mean unfolded structure. SS,  $R_{\text{gyr}}$  and SASA refer to the same observables as described in Table 1.

from the mean equilibrated simulated native structure $\dagger$ ,  $\langle \text{dRMS} \rangle$ , over time: we calculate  $\text{C}^\alpha$ -dRMS from the mean native matrix for each individual unfolded molecule at a given time point, and then average over the entire ensemble at that time point. The bottom curves, on the other hand, give at each time point the  $\text{C}^\alpha$ -dRMS between the mean native matrix and the mean unfolded matrix revealing our central result. The

$\dagger$  The representative time points (see Figures and Tables) for calculating the mean matrices for the unfolded and the native ensembles were chosen to illustrate the main point of the report. The conclusions in no way depend on the exact choice of the time point, as can be seen in the time course figures. For the native ensembles we picked points at which we believe sufficient sampling has been achieved without compromising the structural integrity of the native configuration (Table 1). For the unfolded ensembles we picked points at which the separation between the dRMS of the mean structure and the ensemble-averaged  $\langle \text{dRMS} \rangle$  is greatest. Note that there are points along the time course at which the dRMS of the mean structure actually reaches lower absolute values than the ones at the representative time point. However, as can be seen in the time course figures (Figures 1(c), 2(c) and 3(c)), these differences are insignificant.

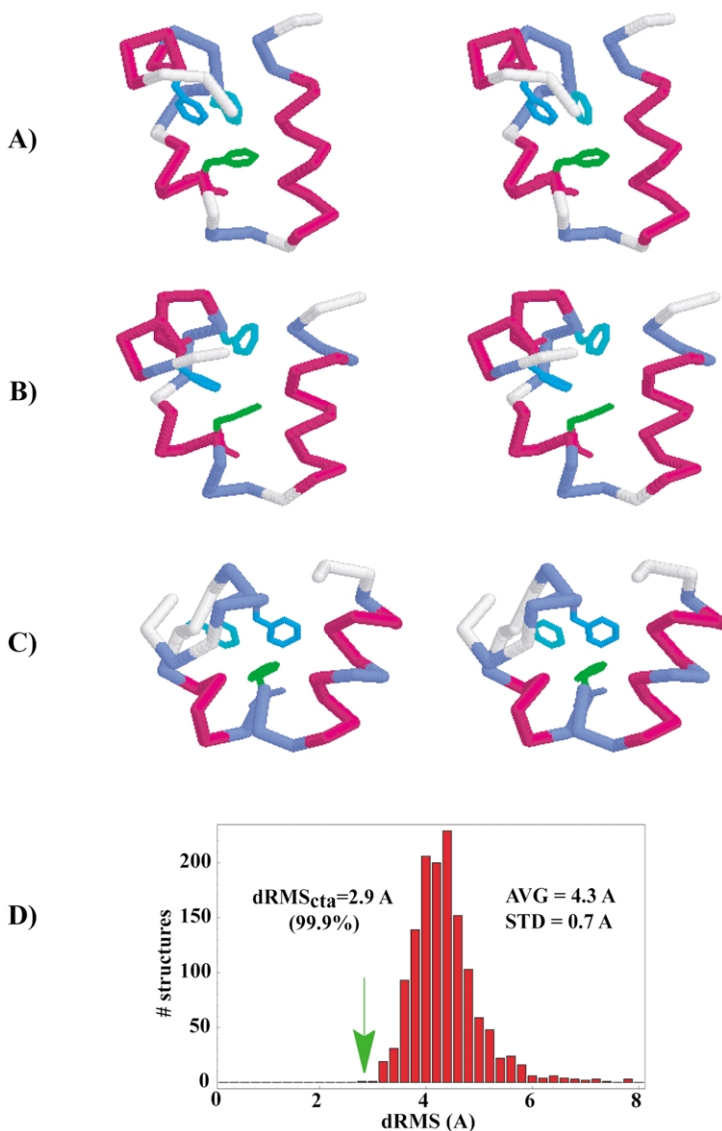


**Figure 4.** Time-averaged results. We compare the 1  $\mu$ s long Duan & Kollman villin trajectory<sup>25</sup> with our native villin ensemble at 20 ns. The distribution of individual dRMS from the mean native structure in our simulations at 20 ns for each of the 50,000 structures spaced by 20 ps is shown. The arrow marks the dRMS of the mean structure on the basis of all 50,000 Duan & Kollman structures from the mean native structure in our simulations at 20 ns. All dRMS values refer to  $\text{C}^\alpha$ -dRMS. The mean (AVG) and the standard deviation (STD) of the depicted distribution is shown. The percentage refers to the fraction of individual structures that are more distant in the dRMS sense from the mean native matrix than is the mean unfolded matrix.

mean unfolded matrix at any given time point is significantly more similar to the mean native matrix than the vast majority of individual unfolded structures!

This effect is further illustrated when we compare the  $\text{C}^\alpha$ -dRMS distributions for individual unfolded molecules with the  $\text{C}^\alpha$ -dRMS of the mean unfolded matrix (all with respect to the mean native matrix) at select, representative time points (Figures 1(d), 2(d), and 3(d)). In the case of villin, for example, the mean unfolded matrix at 27 ns is more similar to the mean native matrix than any of the 5248 individual unfolded structures found at that time point (Figure 1(d)). Furthermore, the  $\text{C}^\alpha$ -dRMS between the mean unfolded matrix and the mean native matrix for all three molecules gets so low (2.3 Å, 2.5 Å, and 2.0 Å for villin, TrpZip and BBA5, respectively) that it suggests the following hypothesis: the geometry of the mean unfolded state in these molecules is close to the mean native geometry, despite the fact that most individual structures are significantly less native-like (e.g. have a greater  $\text{C}^\alpha$ -dRMS).

Finally, in addition to comparing the unfolded ensembles with the simulated native ensembles, we have also compared our unfolded state ensembles to the experimental NMR refined structures. We find similar results, with the only difference being that the distributions and average dRMS values were shifted to higher values by about 0.5–1 Å (Table 2). We have also carried out the above comparisons for the  $\text{C}^\beta$  based distance matrices, and the results show identical trends (Table 2). This indicates that the mean structure assumes not only the native-like backbone



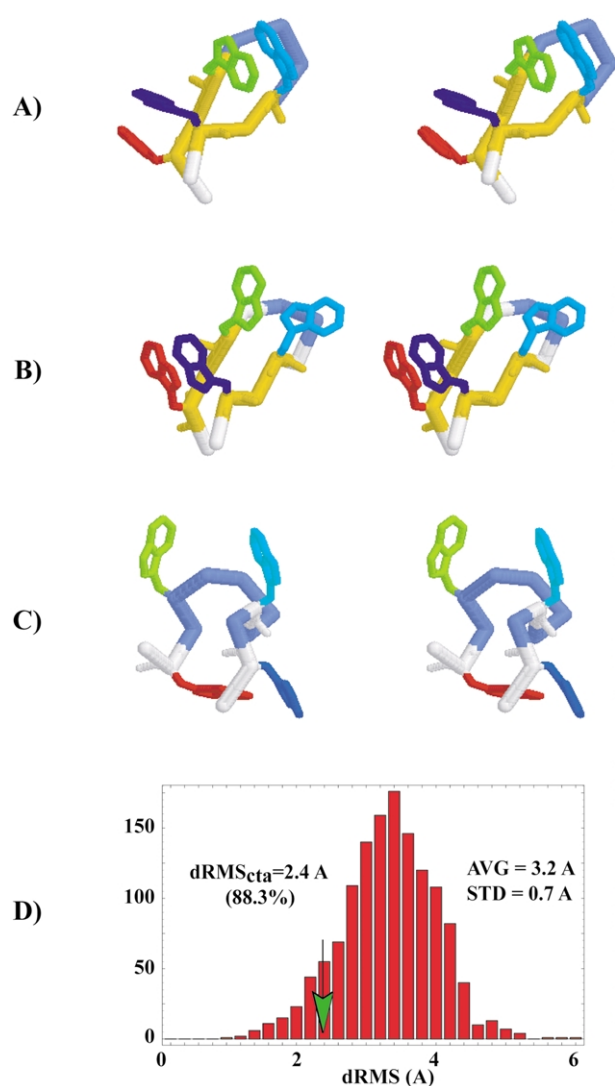
**Figure 5.** Mean structure hypothesis and structure prediction: villin simulations. (a) Stereo image of the experimental NMR structure of villin. (b) Stereo image of the representative structure from our native villin simulations: this structure is the closest (in the  $C^\alpha$ -dRMS sense) individual structure in our simulated folded ensemble at 20 ns to the mean distance matrix on the basis of the same ensemble. The  $C^\alpha$ -dRMS between this structure and the experimental structure in (a) is 1.7 Å. The backbone RMSD between the two structures for all residues is 2.3 Å. (c) Stereo image of the representative structure from our simulations of the unfolded ensemble of villin at 42 ns: this structure is the closest (in the  $C^\alpha$ -dRMS sense) individual structure in our simulated unfolded ensemble at 42 ns to the mean distance matrix on the basis of the same ensemble. The  $C^\alpha$ -dRMS between this structure and the mean native matrix is 2.9 Å. The  $C^\alpha$ -dRMS between this structure and the representative simulated native structure in (b) is 3.3 Å. The backbone RMSD between the two structures for all residues is 4.2 Å. In (a)–(c) the phenylalanine core residues (Phe7, Phe11, Phe18) are shown in color. The secondary structure content is color-coded on the basis of DSSP as follows: red, alpha helix; blue, turn; white, random coil. (d) The green arrow marks the  $C^\alpha$ -dRMS of the representative structure in (c) from the mean native matrix ( $dRMS_{cta}$ , where cta is “closest to average”).

We compare it with the distribution of  $C^\alpha$ -dRMS from the same mean native matrix for all other members of the unfolded ensemble at 42 ns. We show the mean (AVG) and the standard deviation (STD) of the depicted distribution, as well as the percentage (%) of the individual unfolded structures that are more different from the mean native matrix in the dRMS sense than the closest-to-average unfolded structure.

geometry, but also near native side-chain packing, all while individual molecules remain largely unstructured.

The above results pertain to ensemble averages over many short trajectories. It is interesting to speculate whether the same would hold if one averaged over a single long trajectory. We have carried out the above calculations for the 1  $\mu$ s villin folding trajectory simulated by Duan & Kollman,<sup>25</sup> and determined the mean distance matrix over all structures in that simulation (a total of 50,000 structures, sampled every 20 ps, from their 1  $\mu$ s simulation; Figure 4). While it is not obvious to what degree the Duan–Kollman simulation captures the unfolded state, as opposed to some non-native intermediate state, it is clear that the majority of the structures from these simulations do not have the native features.<sup>25</sup> When compared

with the average native matrix at 20 ns from our simulations, it is apparent that the average matrix on the basis of the entire Duan–Kollman trajectory is much closer to the equilibrium mean native matrix than most individual structures (>99.99%), as was seen in the case of ensemble-averaged data (Figure 4). We have also compared the  $C^\beta$ -dRMS values for the Duan–Kollman data:  $C^\beta$ - $\langle dRMS \rangle = 5.3(\pm 0.7)$  Å;  $C^\beta$ - $dRMS_{ms} = 3.5$  Å (better than 100% individual). Finally, we have compared the unfolded data from the Duan–Kollman simulations with their single 100 ns long simulation of the native state. While the average unfolded matrix is still significantly more similar to the native matrix than most individual unfolded structures (>98%), all of values are shifted to higher dRMS values by about 0.5–1 Å ( $C^\alpha$ - $\langle dRMS \rangle = 4.0$  Å,  $C^\alpha$ - $\langle dRMS \rangle = 5.5(\pm 1.0)$  Å). This can probably be



**Figure 6.** Mean structure hypothesis and structure prediction: tryptophan zipper simulations. (a) Stereo image of the experimental NMR structure of tryptophan zipper. (b) Stereo image of the representative structure from our native tryptophan zipper simulations: this structure is the closest (in the  $C^\alpha$ -dRMS sense) individual structure in our simulated folded ensemble at 15 ns to the mean distance matrix on the basis of the same ensemble. The  $C^\alpha$ -dRMS between this structure and the experimental structure in (a) is 0.5 Å. The backbone RMSD between the two structures for all residues is 1.2 Å. (c) Stereo image of the representative structure from our simulations of the unfolded ensemble of tryptophan zipper at 14 ns: this structure is the closest (in the  $C^\alpha$ -dRMS sense) individual structure in our simulated unfolded ensemble at 14 ns to the mean distance matrix on the basis of the same ensemble. The  $C^\alpha$ -dRMS between this structure and the mean native matrix is 2.4 Å. The  $C^\alpha$ -dRMS between this structure and the representative simulated native structure in (b) is 2.6 Å. The backbone RMSD between the two structures for all residues is 3.6 Å. In (a)–(c), the tryptophan core residues (Trp2, Trp4, Trp9, and Trp11) are shown in color. The secondary structure content is color-coded on the basis of DSSP as follows: yellow, beta-sheet; blue, turn; white, random coil. (d) The green arrow marks the  $C^\alpha$ -dRMS of the representative structure in (c) from the mean native matrix ( $dRMS_{cta}$ , where cta is closest to average). We

attributed to the limited extent of their sampling of the native state.

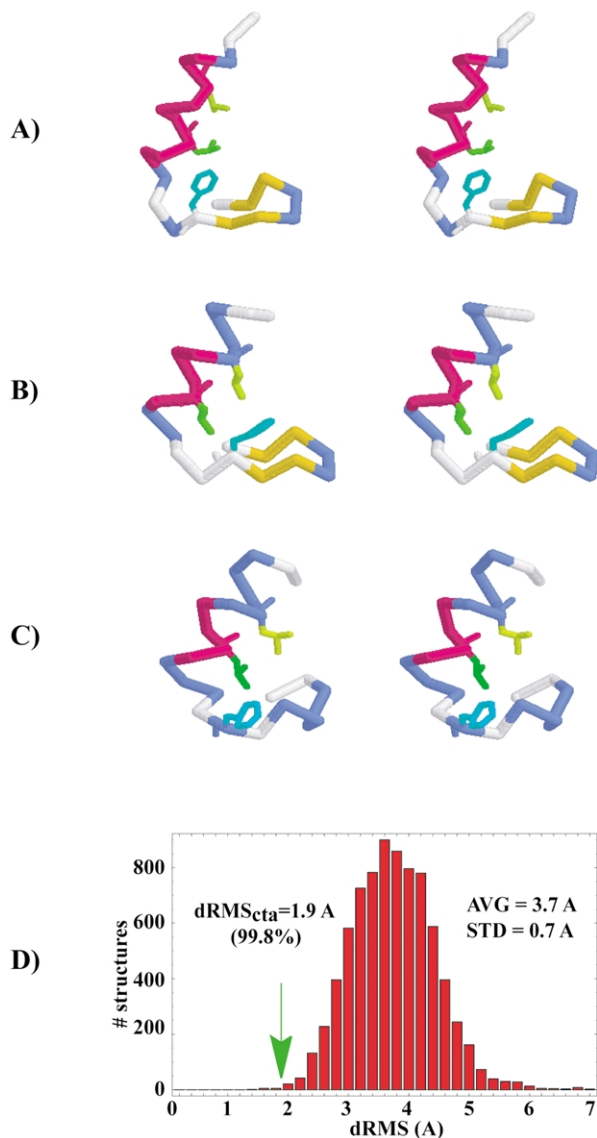
The fact that all three structural categories (alpha, alpha/beta, and beta) gave similar results suggests that the observed phenomenon may be quite general. It is possible, however, that the conclusions pertain only to small proteins and peptides. We are currently simulating several larger proteins to test this idea. Shortle & Ackerman have recently shown that a 131-residue fragment of staphylococcal nuclease retains the overall features of its native topology even when “fully” denatured in 8 M urea.<sup>10</sup> Our results suggest that this is possible even if none of the individual molecules are folded at any given time. The key requirement is that on average the backbone and the side-chains are placed in their native locations. Steric repulsion in the collapsed state may be of critical importance in ensuring that these fluctuations revolve around the native topology.<sup>10,27–29</sup> Finally, the staphylococcal nuclease results suggest that the phenomena reported here may hold even for larger single domain proteins and in the presence of denaturants. It is possible, however, that the denatured staphylococcal nuclease resembles the native staphylococcal nuclease for reasons unrelated to conformational averaging seen in our simulations. Further work will be needed to fully address this question.

## Discussion and Implications

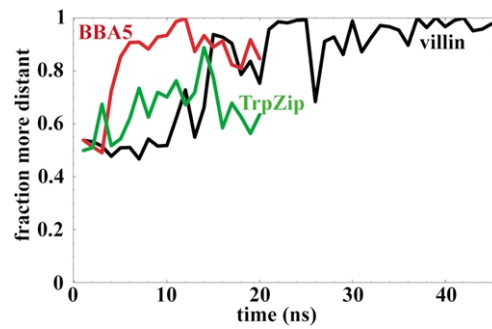
Our findings above lead us to form what we call the “mean-structure hypothesis”, i.e. that the geometry of the collapsed unfolded state of small peptides and proteins in an average sense corresponds to the geometry of the native equilibrium state. Below, we outline several key implications of this hypothesis for the folding of small proteins from both fundamental biophysical as well as methodological perspectives.

First, the results presented above suggest a picture of the folding process in which the structure of the protein in an average sense essentially does not change throughout folding, and it corresponds to the final equilibrium structure. The mean structure stays in place, while folding involves reducing the structural variability of the ensemble. It is as if the distribution of a protein’s structure in some large-dimensional structural space remains centered around the same mean

compare it with the distribution of  $C^\alpha$ -dRMS from the same mean native matrix for all other members of the unfolded ensemble at 14 ns. We show the mean (AVG) and the standard deviation (STD) of the depicted distribution, as well as the percentage (%) of the individual unfolded structures that are more different from the mean native matrix in the dRMS sense than the closest-to-average unfolded structure.



**Figure 7.** Mean-structure hypothesis and structure prediction: BBA5 simulations. (a) Stereo image of the experimental NMR structure of BBA5. (b) Stereo image of the representative structure from our BBA5 simulations: this structure is the closest (in the  $C^\alpha$ -dRMS sense) individual structure in our simulated folded ensemble at 15 ns to the mean distance matrix on the basis of the same ensemble. The  $C^\alpha$ -dRMS between this structure and the experimental structure in (a) is 2.3 Å. The backbone RMSD between the two structures for all residues is 2.9 Å. (c) Stereo image of the representative structure from our simulations of the unfolded ensemble of BBA5 at 12 ns: this structure is the closest (in the  $C^\alpha$ -dRMS sense) individual structure in our simulated unfolded ensemble at 12 ns to the mean native matrix on the basis of the same ensemble. The  $C^\alpha$ -dRMS between this structure and the mean native matrix is 1.9 Å. The  $C^\alpha$ -dRMS between this structure and the representative simulated native structure in (b) is 2.0 Å. The backbone RMSD between the two structures for all residues is 2.3 Å. In (a)–(c) the hydrophobic core residues (Phe9, Leu15, Leu19) are shown in color. The secondary structure content is color-coded on the basis of DSSP as follows: red, alpha-helix; yellow, beta-sheet, blue, turn; white, random coil. (d) The green arrow marks the  $C^\alpha$ -dRMS of the representative structure in (c) from the



**Figure 8.** The filtering power of the mean structure hypothesis. At each time point, we repeat the same calculation as described for Figures 5(c), 6(c), and 7(c) for all three molecules studied: at each time point we find the closest unfolded structure in the dRMS sense to the mean unfolded matrix at that time, and calculate what fraction of the unfolded ensemble at that time is more distant in terms of dRMS from the mean native matrix than this structure. As before, we use the mean native structures at 20, 15, and 15 ns for villin, TrpZip and BBA5, respectively. All dRMS values refer to  $C^\alpha$ -dRMS.

throughout folding, yet its variance narrows as one gets closer to the equilibrium state. It is important to emphasize that this picture describes what happens with the protein on average: the folding pathways of individual proteins may differ significantly from each other as well as from the average pathway, as has been suggested before. Finally, our picture allows for cooperative transitions along any individual trajectory: when any individual protein attains its native shape (and this may happen in an all-or-none, cooperative fashion), this event preserves the mean of the structural distribution, but decreases its variance. A similar overall picture of folding has been suggested by Shortle's group on the basis of their study of the staphylococcal nuclease.<sup>27,28</sup>

Second, it is interesting to consider the implications of the mean structure hypothesis for ensemble *versus* single molecule experiments. It is clear that the average behavior of an ensemble of protein structures can exhibit properties that are not present in any single structure or single trajectory, especially if it is short relative to the relevant characteristic times involved. Since typical experiments (e.g. NMR and X-ray) “look” at structures in an average sense, we believe that the most informative way of comparing any simulation

mean native matrix ( $dRMS_{cta}$ , where cta is closest to average). We compare it with the distribution of  $C^\alpha$ -dRMS from the same mean native matrix for all other members of the unfolded ensemble at 12 ns. We show the mean (AVG) and the standard deviation (STD) of the depicted distribution, as well as the percentage (%) of the individual unfolded structures that are more different from the mean native matrix in the dRMS sense than the closest-to-average unfolded structure.

with experiment should involve simulating large ensembles of proteins, averaging the structural information in the same way as the experiment, and then comparing the results directly with experimental data (NOE distance restraints or structure factors).<sup>9,30–33</sup> The average distance matrices in the present study are in spirit akin to the structural information coming from NMR experiments. We are currently calculating theoretical NOEs using a full matrix relaxation approach to even more closely match the experiment (unpublished results). Recently, a computational study of short  $\beta$ -peptides in methanol was reported<sup>33</sup> in which it was shown that the ensemble averaged NOEs and  $J$ -coupling constants are quite insensitive to the degree of diversity of the underlying ensemble.

Moreover, since the average of an ensemble may hide certain properties, in particular its heterogeneity, we believe that one must examine the variance of structural distributions in addition to the mean. For example, two structures may look the same on average, yet one may fluctuate around this average much more than the other: depending on the experimental technique, it may appear as if one is formed and the other one is not. Accordingly, an ensemble may look structured on average, while no single member is structured. For instance, as we propose, the collapsed unfolded state has the mean features of the native state, but is characterized by high structural variance. Single molecule analysis, with its ability to access the variance or distribution within the behavior of individual species, may prove invaluable in addressing these timely questions.

Third, the mean-structure hypothesis suggests a way of performing and/or supplementing protein structure prediction using the information gleaned from the simulation of a large unfolded ensemble very early into folding. Using the distance constraints derived from the average structure of the unfolded state, one could in principle, refine a structure, and if the mean-structure hypothesis is correct, this structure should be native-like. Alternatively, one could find the closest (in the dRMS sense) individual member of the unfolded ensemble to the average structure based on the same unfolded ensemble; if the hypothesis is correct, this structure should be closer to the native structure than most other individual unfolded structures.

Analysis of our data set in this manner strongly supports this argument. The comparison of a representative native villin structure (Figure 5(b)) with the member of the unfolded ensemble at 42 ns that is closest to the mean unfolded structure at that time (Figure 5(c)) shows that the mean of the unfolded ensemble has the native topology, most of native secondary structure and reasonable core packing. In fact, the structure closest to the mean is closer to the structure of the native state in the dRMS sense than 99.9% of the 1211 unfolded structures sampled at 42 ns (Figure 5(d)). Further-

more, even though most individual molecules in the unfolded ensemble exhibit very little helical secondary structure (see Figure 1(b) inset and Table 2), the closest-to-average structure shown in Figure 5(c) has almost native-like helical content.

The same analysis for tryptophan zipper and BBA5 is shown in Figures 6 and 7, and the results are as encouraging for BBA5 and less so in the case of tryptophan zipper (see below for a discussion of the tryptophan zipper results). The representative unfolded structures in Figures 5(c), 6(c) and 7(c) can be thought of as predictions of the native state in a scheme where the unfolded ensemble plays a role of a decoy set and the dRMS distance from the mean matrix on the basis of the entire decoy set (i.e. the entire unfolded ensemble at a given time point) serves as a simple “energy function” or scoring function for decoy discrimination. The filtering power of this scoring function is demonstrated in Figures 5(d), 6(d) and 7(d): the unfolded structure closest to the mean unfolded matrix is closer to the mean native matrix than the majority of individual unfolded structures in all three cases. What is more, this fact holds for the majority of the time points after the collapse in villin and BBA5 simulations, and to a lesser degree in TrpZip simulations (Figure 8). Note that structures in Figures 5(c), 6(c) and 7(c) were picked with no knowledge of the native structure, and can be thought of as blind predictions for the given time points.

These results and arguments greatly extend and support previous theoretical findings. Shortle *et al.*<sup>34</sup> have observed in their *ab initio* prediction studies that clusters with the largest number of similar structures in their decoy sets are likely to contain the native structure. Our unfolded ensembles are akin to decoy sets in a sense that they contain many low energy structures, which, to a varying degree, all differ from the native state. It is clear that the average structure of the unfolded ensemble will be dominated by the most represented structure or set of structures, and in this way our results are compatible with the findings by Shortle *et al.* Finally, Huang and colleagues<sup>35,36</sup> have shown that finding the average structure from a set of decoys on the basis of the most represented  $C^\alpha$ – $C^\alpha$  distances yields structures that are closer to the native structure than most individual decoy structures. Our findings suggest that these results are a consequence of certain intrinsic properties of polypeptides. To use the above approach in structure prediction, however, very large ensembles are necessary: likely large enough to see folding in the timescale simulated. With the advent of distributed computing, such calculations are now well within reach, and can be performed on the scale of days on large (10,000–100,000) processor clusters. It is also possible that the principles outlined above can be applied to ensembles generated with cheaper simulation methods (such as Monte Carlo simulations).

We have seen that the  $\beta$ -sheet tryptophan zipper molecule, while still significant, is less supportive of the mean structure hypothesis than villin and BBA5. This is apparent for both the averaged matrices from the unfolded ensemble (Figure 3), and even more so, for the closest real unfolded structure to the mean unfolded matrix (Figure 6(c)). Except for the overall U-shape, this structure (Figure 6(c)) exhibits neither beta secondary content nor proper core packing. There are several potential reasons for this. The most important reason is probably the specific geometry of the beta hairpin: namely, in a hairpin, the distances between the residues on the same strand are typically as large as can be, and the distances between the residues on the opposing sides as low as can possibly be. What this means is that any kind of linear averaging of distances over a large heterogeneous population will result in values that are smaller than the typical beta intra-strand distances or larger than the typical beta inter-strand distances. This is exactly what we see in our simulations of the tryptophan zipper (Figure 6(c)). The same effect is seen in the case of the representative unfolded structure of BBA5 (Figure 7(c)), where the N-terminal beta-sheet is not formed and the strands are significantly more separated than in a beta-sheet. It is possible that a different kind of averaging (such as  $\langle r^{-6} \rangle^{-1/6}$  averaging) might give different results, and we are currently exploring this possibility. Tryptophan zipper is a small peptide and it is possible that the mean structure hypothesis is more applicable to larger, more globular molecules with real tertiary structure. To test this possibility, we are currently running simulations of a larger, all-beta Trp-Trp domain. Finally, the TrpZip domain is so small that upon collapse it is immediately by necessity already very close in the  $\langle \text{dRMS} \rangle$  sense to the native topology. This naturally precludes the average structure to be any more similar. Finally, it is also possible that, the native state of TrpZip being very stable, we have not sampled enough of it to generate true equilibrium distribution of native structures.

What is the physical and geometric meaning of the mean structure? The average matrices used in the above analysis do not directly correspond to any individual real physical structures. Indeed, they even violate the elementary triangle inequality for certain triplets of distances. This is not surprising if one understands that these matrices are generated on the basis of averaging the information coming from thousands of individual structures. However, in this way they are closely analogous to the real observables coming from ensemble-averaged experiments. For instance, the NOE maps in NMR experiments or the diffraction patterns in X-ray experiments are also based on the average properties of large ensembles of molecules, and furthermore, these observables also rarely perfectly match any individual real structures. The non-zero  $R$  factors in

X-ray refinement or the non-zero  $R_x$  factors in NMR refinement attest to this fact. The results on the basis of averaged matrices in this work should be taken as a statement about something akin to the real data, rather than idealized refined structures. In order to get even closer to the real experimental observables, we are currently calculating the theoretical NOEs on the basis of our simulations using the full relaxation matrix approach and comparing the unfolded and the folded ensembles in that way (unpublished results).

Nevertheless, the results presented in Figures 5–7 show that the main conclusions of this work hold even for real physical structures. By finding the closest individual member of the unfolded ensemble in the  $C^\alpha$ -dRMS sense to the mean distance matrix based on the same ensemble, we have in a way performed refinement. That the structures refined in such a way bear close resemblance to the representative simulated native structures in the case of villin and BBA5, and somewhat less so in the case of tryptophan zipper, offers strong support to the claim that the averaged unfolded distance matrices are indeed close to the physical native structures.

The mean structure hypothesis is a statement about the effects of conformational averaging in the collapsed, unfolded ensembles of small proteins. On the basis of the results presented here, it appears applicable to globular, mostly helical proteins. Further work is needed to elucidate its range of validity in other systems, most notably larger globular proteins.

## Methods

Using a heterogeneous computer cluster we have generated thousands (see the text, Figures 1(a), 2(a) and 3(a)) of relatively short (tens of nanoseconds) independent trajectories for the villin, tryptophan zipper, and BBA5 molecules. With the exception of the data produced by Duan & Kollman,<sup>25</sup> the simulations described below were generated with the following methodology. The folding simulations were initiated from fully extended conformations ( $\phi = -135^\circ$ ,  $\psi = 135^\circ$ ) with *N*-acetyl and *C*-amino caps. The equilibrium simulations were started from the experimental NMR structures of the three molecules (villin,<sup>22</sup> PDB-code 1VII, first structure, sequence: MLSEDFKAVFGMTRSAFANL-PLWKQQNLKKEKGLF; tryptophan-zipper PDB-code 1HRW,<sup>23</sup> first structure, sequence: SWTWEGNKWTWK†; BBA5,<sup>24</sup> sequence: YRVPSYDFSRSEDELAKLLRQHAG).

† After submission of this manuscript, we have found out that the experimental structure of the tryptophan zipper (1HRW) has been recalled and replaced in the RCSB protein data bank with an improved highly similar structure (1LE0). The main difference concerns the  $\chi_1$  angles of the two outer tryptophan residues (Trp9 and 11): they have been changed from  $-60$  to  $180$  degrees. The backbone rmsd between the old structure used in this study and the newly reported one is  $0.51 \text{ \AA}$ . This difference should not have any significant effect on any of the conclusions presented here.

Villin headpiece is a 36 residue three-helix bundle protein; tryptophan zipper is a 12 residue  $\beta$ -hairpin peptide with a core of four tryptophan residues; BBA5 is a 23 residue designed mini-protein consisting of an  $\alpha$ -helix and a  $\beta$ -hairpin packed perpendicularly to each other. All three molecules are expected to fold extremely rapidly (in the tens of microseconds regime). The simulations, run using Tinker biomolecular simulation package $\ddagger$ , involved Langevin dynamics in implicit generalized Born/surface area (GB/SA) $^{37}$  solvent (viscosity of 91 ps $^{-1}$ ) with a 2 fs integration step, at temperatures indicated in Tables 1 and 2. Bond lengths were constrained using RATTLE. $^{38}$  16 Å cutoffs with 12 Å tapers were used for electrostatics in TrpZip and BBA5 simulations. No cutoffs were used for villin. The proteins were modeled using the OPLSua force field. $^{39}$  The structures were output for analysis every 1 ns of simulated time. The simulations were carried out on 10,000+ processors as a part of our ongoing Folding@Home distributed computing project $\S$ , and involved a total of about half a trillion ( $4 \times 10^{11}$ ) integration steps. This corresponds to approximately 2000 single CPU (500 MHz) years.

## Acknowledgements

We thank all the contributors to the Folding@Home project: a complete list can be found at <http://folding.stanford.edu>. B.Z. and C.D.S. acknowledge support from the HHMI Predoctoral Fellowship program. We thank Robert Baldwin, Kevin Plaxco, Michael Levitt, Richard Aldrich, Sebastian Doniach, Stefan Larson and Eric Sorin for useful comments, and Yong Duan for giving us access to his data. This work was supported by grants from the ACS PRF (36028-AC4), NIH BBCA (R01GM62868-01A2), NSF MRSEC CPIMA (DMR-9808677), NIH BISTI (IP20 GM64782-01), ARO (41778-LS-RIP), and Stanford University (Internet 2), as well as by gifts from the Intel and Google corporations. B.Z. dedicates this work to his grandmother Slavica.

## References

1. Plaxco, K. W. & Gross, M. (2001). Unfolded, yes, but random? Never! *Nature Struct. Biol.* **8**, 659–660.
2. Ptitsyn, O. B. (1995). Molten globule and protein folding. *Advan. Protein Chem.* **47**, 83–229.
3. Zhang, O., Kay, L. E., Shortle, D. & Forman-Kay, J. D. (1997). Comprehensive NOE characterization of a partially folded large fragment of staphylococcal nuclease delta131delta, using nmr methods with improved resolution. *J. Mol. Biol.* **272**, 9–20.
4. Bond, C. J., Wong, K. B., Clarke, J., Fersht, A. R. & Daggett, V. (1997). Characterization of residual structure in the thermally denatured state of barnase by simulation and experiment: description of the folding pathway. *Proc. Natl Acad. Sci. USA*, **94**, 13409–13413.
5. Srinivasan, R. & Rose, G. D. (1999). A physical basis for protein secondary structure. *Proc. Natl Acad. Sci.* **96**, 14258–14263.
6. Alonso, D. O. & Daggett, V. (2000). Staphylococcal protein A: unfolding pathways, unfolded states, and differences between the B and E domains. *Proc. Natl Acad. Sci. USA*, **97**, 133–138.
7. Mayor, U., Johnson, C. M., Daggett, V. & Fersht, A. R. (2000). Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation. *Proc. Natl Acad. Sci. USA*, **97**, 13518–13522.
8. van Gunsteren, W. F., Burgi, R., Peter, C. & Daura, X. (2001). 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.
9. Kazmirski, S. L. & Daggett, V. (1998). Simulations of the structural and dynamical properties of denatured proteins: the “molten coil” state of bovine pancreatic trypsin inhibitor. *J. Mol. Biol.* **277**, 487–506.
10. Shortle, D. & Ackerman, M. S. (2001). Persistence of native-like topology in a denatured protein in 8 M urea. *Science*, **293**, 487–489.
11. Shi, Z., Olson, C. A., Rose, G. D., Baldwin, R. L. & Kallenbach, N. R. (2002). Polyproline II structure in a sequence of seven alanine residues. *Proc. Natl Acad. Sci. USA*, **99**, 9190–9195.
12. Fersht, A. R. & Daggett, V. (2002). Protein folding and unfolding at atomic resolution. *Cell*, **108**, 573–582.
13. Lecomte, J. T. & Falzone, C. J. (1999). Where U and I meet. *Nature Struct. Biol.* **6**, 605–608.
14. Bai, Y., Chung, J., Dyson, H. J. & Wright, P. E. (2001). Structural and dynamic characterization of an unfolded state of poplar apo-plastocyanin formed under non-denaturing conditions. *Protein Sci.* **10**, 1056–1066.
15. Mok, Y. K., Kay, C. M., Kay, L. E. & Forman-Kay, J. (1999). NOE data demonstrating a compact unfolded state for an SH3 domain under non-denaturing conditions. *J. Mol. Biol.* **289**, 619–638.
16. Tollinger, M., Skrynnikov, N. R., Mulder, F. A., Forman-Kay, J. D. & Kay, L. E. (2001). Slow dynamics in folded and unfolded states of an SH3 domain. *J. Am. Chem. Soc.* **123**, 11341–11352.
17. Li, A. & Daggett, V. (1996). Identification and characterization of the unfolding transition state of chymotrypsin inhibitor 2 by molecular dynamics simulations. *J. Mol. Biol.* **257**, 412–429.
18. Li, A. & Daggett, V. (1998). Molecular dynamics simulation of the unfolding of barnase: characterization of the major intermediate. *J. Mol. Biol.* **275**, 677–694.
19. Kazmirski, S. L., Wong, K. B., Freund, S. M., Tan, Y. J., Fersht, A. R. & Daggett, V. (2001). Protein folding from a highly disordered denatured state: the folding pathway of chymotrypsin inhibitor 2 at atomic resolution. *Proc. Natl Acad. Sci. USA*, **98**, 4349–4354.
20. Shirts, M. R. & Pande, V. S. (2001). Mathematical analysis of coupled parallel simulations. *Phys. Rev. Letters*, **86**, 4983–4987.
21. Zagrovic, B., Sorin, E. J. & Pande, V. (2001). Beta-hairpin folding simulations in atomistic detail using an implicit solvent model. *J. Mol. Biol.* **313**, 151–169.
22. McKnight, C. J., Matsudaira, P. T. & Kim, P. S. (1997). NMR structure of the 35-residue villin headpiece subdomain. *Nature Struct. Biol.* **4**, 180–184.
23. Cochran, A. G., Skelton, N. J. & Starovasnik, M. A. (2001). Tryptophan zippers: stable, monomeric beta-hairpins. *Proc. Natl Acad. Sci. USA*, **98**, 5578–5583.

$\ddagger$  <http://dasher.wustl.edu/tinker/>

$\S$  <http://folding.stanford.edu>

24. Struthers, M., Ottesen, J. J. & Imperiali, B. (1998). Design and NMR analyses of compact, independently folded BBA motifs. *Fold Des.* **3**, 95–103.
25. Duan, Y. & Kollman, P. A. (1998). Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, **282**, 740–744.
26. Ferrara, P. & Caffisch, A. (2000). Folding simulations of a three-stranded antiparallel beta-sheet peptide. *Proc. Natl Acad. Sci. USA*, **97**, 10780–10785.
27. Gillespie, J. R. & Shortle, D. (1997). 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.
28. Gillespie, J. R. & Shortle, D. (1997). 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.
29. Pappu, R. V., Srinivasan, R. & Rose, G. D. (2000). The Flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding. *Proc. Natl Acad. Sci. USA*, **97**, 12565–12570.
30. Shortle, D., Wang, Y., Gillespie, J. R. & Wrabl, J. O. (1996). Protein folding for realists: a timeless phenomenon. *Protein Sci.* **5**, 991–1000.
31. Burgi, R., Pitera, J. & van Gunsteren, W. F. (2001). Assessing the effect of conformational averaging on the measured values of observables. *J. Biomol. NMR*, **19**, 305–320.
32. Krueger, B. P. & Kollman, P. A. (2001). Molecular dynamics simulations of a highly charged peptide from an SH3 domain: possible sequence-function relationship. *Proteins: Struct. Funct. Genet.* **45**, 4–15.
33. Daura, X., Antes, I., van Gunsteren, W. F., Thiel, W. & Mark, A. E. (1999). The effect of motional averaging on the calculation of NMR-derived structural properties. *Proteins: Struct. Funct. Genet.*, **36**, 542–555.
34. Shortle, D., Simons, K. T. & Baker, D. (1998). Clustering of low-energy conformations near the native structures of small proteins. *Proc. Natl Acad. Sci. USA*, **95**, 11158–11162.
35. Huang, E. S., Samudrala, R. & Ponder, J. W. (1998). Distance geometry generates native-like folds for small helical proteins using the consensus distances of predicted protein structures. *Protein Sci.* **7**, 1998–2003.
36. Huang, E. S., Samudrala, R. & Ponder, J. W. (1999). *Ab initio* fold prediction of small helical proteins using distance geometry and knowledge-based scoring functions. *J. Mol. Biol.* **290**, 267–281.
37. Qiu, D., Shenkin, P. S., Hollinger, F. P. & Still, W. C. (1997). The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. *J. Phys. Chem.* 3005–3014.
38. Andersen, H. C. (1983). Rattle: a “velocity” version of the shake algorithm for molecular dynamics calculations. *J. Comput. Phys.* **52**, 24–34.
39. Jorgensen, W. L. & Tirado-Rives, J. (1988). The OPLS potential functions for proteins: energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **110**, 1666–1671.
40. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

*Edited by P. E. Wright*

(Received 3 May 2002; received in revised form 9 August 2002; accepted 14 August 2002)