

Simulation of Folding of a Small Alpha-helical Protein in Atomistic Detail using Worldwide-distributed Computing

Bojan Zagrovic¹, Christopher D. Snow¹, Michael R. Shirts² and Vijay S. Pande^{1,2*}

¹*Biophysics Program, Stanford University, Stanford, CA 94305-5080, USA*

²*Department of Chemistry, Stanford University, Stanford CA 94305-5080, USA*

By employing thousands of PCs and new worldwide-distributed computing techniques, we have simulated in atomistic detail the folding of a fast-folding 36-residue α -helical protein from the villin headpiece. The total simulated time exceeds 300 μ s, orders of magnitude more than previous simulations of a molecule of this size. Starting from an extended state, we obtained an ensemble of folded structures, which is on average 1.7 Å and 1.9 Å away from the native state in C $^{\alpha}$ distance-based root-mean-square deviation (dRMS) and C $^{\beta}$ dRMS sense, respectively. The folding mechanism of villin is most consistent with the hydrophobic collapse view of folding: the molecule collapses non-specifically very quickly (\sim 20 ns), which greatly reduces the size of the conformational space that needs to be explored in search of the native state. The conformational search in the collapsed state appears to be rate-limited by the formation of the aromatic core: in a significant fraction of our simulations, the C-terminal phenylalanine residue packs improperly with the rest of the hydrophobic core. We suggest that the breaking of this interaction may be the rate-determining step in the course of folding. On the basis of our simulations we estimate the folding rate of villin to be approximately 5 μ s. By analyzing the average features of the folded ensemble obtained by simulation, we see that the mean folded structure is more similar to the native fold than any individual folded structure. This finding highlights the need for simulating ensembles of molecules and averaging the results in an experiment-like fashion if meaningful comparison between simulation and experiment is to be attempted. Moreover, our results demonstrate that (1) the computational methodology exists to simulate the multi-microsecond regime using distributed computing and (2) that potential sets used to describe interatomic interactions may be sufficiently accurate to reach the folded state, at least for small proteins. We conclude with a comparison between our results and current protein-folding theory.

© 2002 Elsevier Science Ltd. All rights reserved

Keywords: molecular dynamics; protein folding; villin headpiece; ensemble averaging; distributed computing

*Corresponding author

Introduction

Understanding the sequence–structure relationship of proteins will play a pivotal role in the post-genomic era, and may have great impact in genetics, biochemistry, and pharmaceutical chemistry.^{1–3} Other examples of the importance of folding include the understanding of diseases believed to be related to protein-misfolding.⁴ Finally, an understanding of the protein folding mechanism may have great impact on protein

Abbreviations used: dRMS, distance-based root-mean-square deviation; NOE, nuclear Overhauser enhancement; GB/SA, generalized Born/surface area; PMF, potentials of mean force; SASA, solvent-accessible surface area.

E-mail address of the corresponding author: pande@stanford.edu

design as well as on the burgeoning field of nanotechnology, in which self-assembling nanomachines may be designed using synthetic polymers with protein-like folding properties.⁵

Unfortunately, current computational techniques to understand such problems are fundamentally limited by their computational complexity. For example, while the fastest proteins fold on the order of tens of microseconds, current computers can simulate only nanoseconds of real-time folding in full atomic detail in a day, which is a greater than thousand-fold computational gap. A tour-de-force parallelization of simulation code for supercomputers by Duan & Kollman has led to a simulation reaching one microsecond,⁶ but these methods require complex, expensive supercomputers, and still have not led to a simulation of folding (requiring simulation in the tens of microseconds regime).

In addition to direct simulations, several alternative methods have been developed to study protein folding. One of the most powerful such methods, pioneered by Brooks and co-workers,⁷⁻⁹ involves the generation and analysis of the potentials of mean force (PMF). This free energy-based method provides detailed thermodynamic characterization of the system and can, in principle, be used to derive mechanistic information. In addition, the method is trivially parallelizable in the computational sense, and in conjunction with powerful computational resources such as large computer clusters it may play an increasingly important role in the future study of protein folding. Another approach, popularized by Daggett and co-workers, involves high-temperature (e.g. 400–500 K) unfolding simulations.⁹⁻¹³ This method is often used to provide insight into the mechanistic aspects of protein folding. While the regime of applicability of high-temperature simulations to understanding folding under native conditions remains under debate, initial comparison with experiment has provided an important demonstration of their applicability. However, in addition to exploring the behavior of folding at physiological conditions (compared with unfolding at very high temperatures), it is interesting to use a much greater sampling of trajectories (in either folding or unfolding) in order to better assess the mechanism of folding statistically. While both PMFs and high-temperature unfolding provide valuable information about protein behavior and can complement dynamical simulations, many aspects of protein dynamics and folding mechanism can be addressed best only through direct simulation of the kinetics.

Recently, another approach has been developed to bridge the enormous computational gap haunting the protein folding field: worldwide distributed computing.¹⁴⁻¹⁶ There are hundreds of millions of PCs potentially available for calculation, most of which are vastly underused: in fact, these computers could be used to potentially form the most powerful supercomputer on the

planet by several orders of magnitude. However, to tap into this resource efficiently and productively one must employ non-traditional parallelization techniques.^{14,15} Alternatively, one can simply run thousands of independent simulations in parallel. Due to the stochastic nature of folding, one can expect to obtain a small ensemble of complete folding events even if the effective simulated time per simulation is significantly less than the average folding time. In particular, properties of single exponential kinetics can be used to our advantage: proteins that fold on the 10 μ s timescale do not require simulations that are each 10 μ s long. Indeed, since the fraction of simulations that fold is $f(t) = 1 - \exp(-kt)$, for short times, we expect that a fraction kt would fold. For a 10 μ s reaction studied by simulations of length 10 ns, one would expect to find a fraction $f(t = 10 \text{ ns}) = 10 \text{ ns} / 10,000 \text{ ns} = 1/1000$ that fold. While this is too small to expect to get any meaningful results from a single simulation, one would expect to see meaningful results from 10,000 simulations, with which we would expect to see many (ten) successful folding events. In addition, such an approach yields an extremely detailed view of the unfolded ensemble.¹⁷ Here, we demonstrate the application of such a technique using thousands of PCs distributed throughout the world to simulate the folding of a thermostable, fast-folding, 36-residue α -helical subdomain (PDB code 1VII) from the villin headpiece (the C-terminal domain of the much larger villin actin-binding protein).^{6,18}

Results and Discussion

Before we discuss our results, we need to comment briefly on a certain key feature of our analysis. One typically assesses the stability of folded molecules, as well as the overall success of protein simulation in general, in terms of root-mean-square deviation (RMSD) or distance-based root-mean-square deviation (dRMS) of individual structures from the native state. For instance, a single trajectory started from the native state which in, say, 20 ns results in a 5 Å RMSD structure, would be grounds for dismissing the simulation and the force-field as unstable and insufficiently accurate. The same holds for other geometrical descriptors of structure as well.

However, we suggest that one should judge simulations by means directly analogous to experimental measurements. In particular, in typical experiments (e.g. by NMR or X-ray), protein structure and its fluctuations are not measured by examining single molecules, but rather by ensemble measurements. Therefore, we find it imperative to examine ensemble averages of our structural data from simulations before any meaningful comparison with experiment can be contemplated. One way of doing this is by calculating a $C^\alpha-C^\alpha$ (or $C^\beta-C^\beta$) distance matrix for each individual structure in a given ensemble, and then

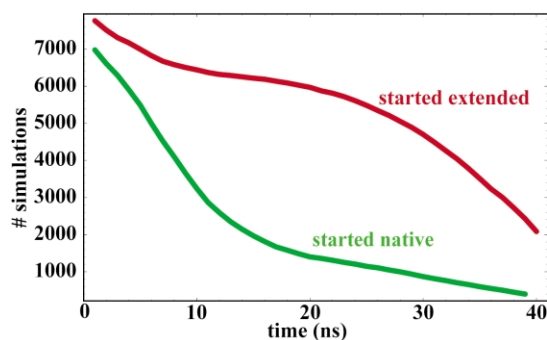


Figure 1. The total number of independent simulations that have reached a given time-point for the equilibrium (“started native”, green) and the folding (“started extended”, red) simulations.

averaging these matrices to obtain one mean matrix representative of the entire ensemble. This approach is, in spirit, very similar to any ensemble-averaged, distance-based structural method such as NMR, electron paramagnetic resonance (EPR), or fluorescence resonance energy transfer (FRET). In this way, for instance, we can look at the average features of an ensemble consisting of all the structures that have folded in our simulations or, say, of an equilibrium ensemble based on simulations started from the experimental structure.

Sampling the native state ensemble

We have performed more than 90 μ s worth of simulation (Figure 1) started from the experimental structure of the villin headpiece (PDB code 1VII).¹⁸ These simulations are important for two reasons. First, before contemplating performing a folding simulation, we must check to ensure that our models (protein force-field, solvation model, etc.) are sufficiently accurate such that the native state is stable in our simulations. By performing long timescale simulations, we can judge the stability of the folded state in our model quantitatively. Second, we need some means to judge whether a simulation has reached the folded state (or at the very least, the folded state of our model). To assess this, we use these simulations as a characterization of the native state ensemble.

We have simulated several thousand fully independent trajectories started from the native structure, each approximately 30 ns long (Figure 1). As shown in Figure 2, the simulations are stable with respect to the compactness (radius of gyration, solvent-accessible surface area (SASA)), secondary structure, and core packing (Figure 2(a) and (b)). The only difference between our simulations and the experimental structure is that, on average, in our simulations we see a slight rotation ($\sim 20^\circ$) of the N-terminal α helix towards the central helix. When we look at the dRMS from the starting

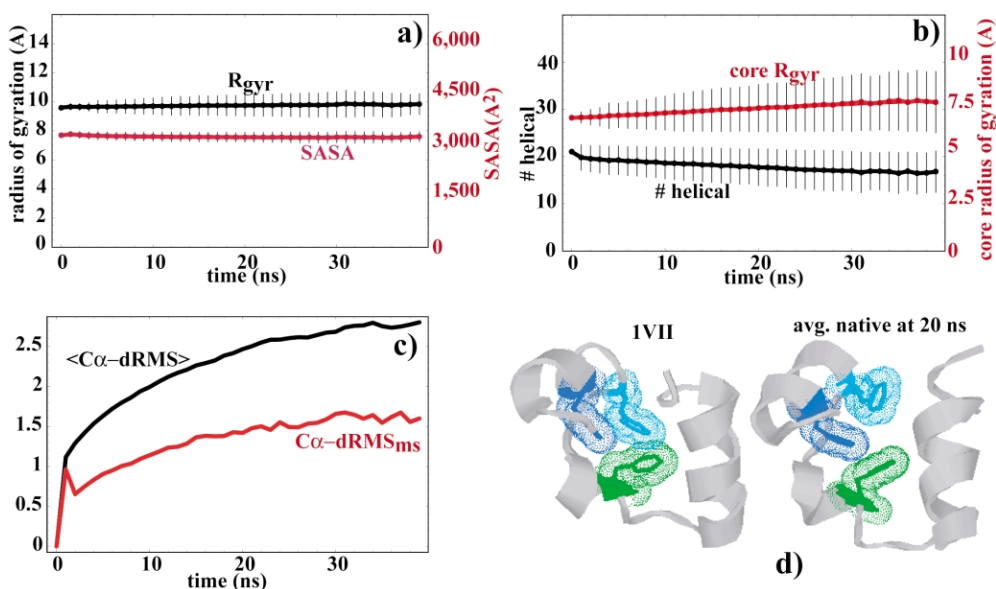


Figure 2. Summary of the equilibrium simulations started from the experimental 1VII structure. (a) Radius of gyration (R_{gyr}), black, and SASA, red, over time for the native simulations. (b) Phenylalanine core radius of gyration (core R_{gyr} —residues 7, 11, 18), black, and the total number of helical residues as determined by DSSP,⁴⁰ red. In both (a) and (b), at each time-point we show the average values calculated over the entire ensemble at that time-point. Error bars represent standard deviation. (c) Comparison over time between the ensemble averaged C^α dRMS, $\langle C^\alpha \text{ dRMS} \rangle$, black, and the C^α dRMS of the mean structure (see the text), $C^\alpha \text{ dRMS}_{\text{ms}}$, red, all with respect to the experimental 1VII structure. (d) Comparison between the experimental 1VII structure,¹⁸ labeled 1VII, and the representative native structure from the simulated ensemble at 20 ns, labeled avg. native at 20 ns. The representative structure is the closest individual structure in the C^α dRMS sense to the mean matrix (see the text) based on the entire native ensemble at 20 ns. The C^α dRMS = 2.0 Å and the main-chain RMSD = 3.0 Å between the two structures (residues 1–36).

structure, and average over the entire native ensemble at different time-points, we see that, on average, the C^α dRMS and C^β dRMS hover around $2.8(\pm 1.0)$ Å (Figure 2(c)) and $3.2(\pm 1.0)$ Å (not shown), respectively. However, when we first average the distance matrices of all the structures at a given time-point, as described above, and then compare this mean distance matrix with the distance matrix based on the experimental structure, we obtain significantly lower values: 1.4 Å (Figure 2(c)) and 1.7 Å (not shown) for C^α dRMS and C^β dRMS, respectively. This latter approach is, in our opinion, more meaningful, since after all, the native structure was derived from an ensemble-averaged distance-based experiment (NMR nuclear Overhauser enhancement (NOE)). When looked at in this fashion, our simulations preserve the integrity of the native state to a high degree.

In Figure 2(d) we compare the starting experimental structure of villin with a member of our simulated native ensemble at 20 ns that was closest in the C^α dRMS sense to the mean C^α distance matrix calculated over the same ensemble at 20 ns. This particular structure can be thought of as the representative structure from our simulated ensemble at 20 ns (which consists of 1405 structures), and its similarity with the original structure in terms of secondary structure, core packing and the overall geometry is quite obvious. Even though some individual structures in our simulated native ensemble unfold and most do exhibit significant fluctuations, our simulations do preserve the integrity of the native structure to a high degree in the mean sense (which is the only relevant sense for comparisons to existing experiments). On the basis of this, we compare our folding ensemble with the mean distance C^α and C^β matrices based on the simulated native ensemble at 20 ns. At this time-point, we believe, sufficient sampling of the native state has been achieved without compromising the structural integrity of the native configuration. The results presented below, however, do not depend significantly on the exact choice of this time-point. Below, we will use the term mean native structure in reference to the averaged distance matrices from native structure simulations.

Folding simulations

We have performed thousands of independent simulations started from an extended structure for tens of nanoseconds each (the total simulated time exceeds 220 μ s, Figure 1). In this way we have captured the folding ensemble very early into folding; our simulations in essence provide a detailed picture of the unfolded state. We start from the extended state to avoid introducing any biases towards the native state. With this great degree of sampling, it is natural to first examine the unfolded state ensemble. Figure 3 details the average geometrical features of the unfolded ensemble over time. The most important feature is probably the

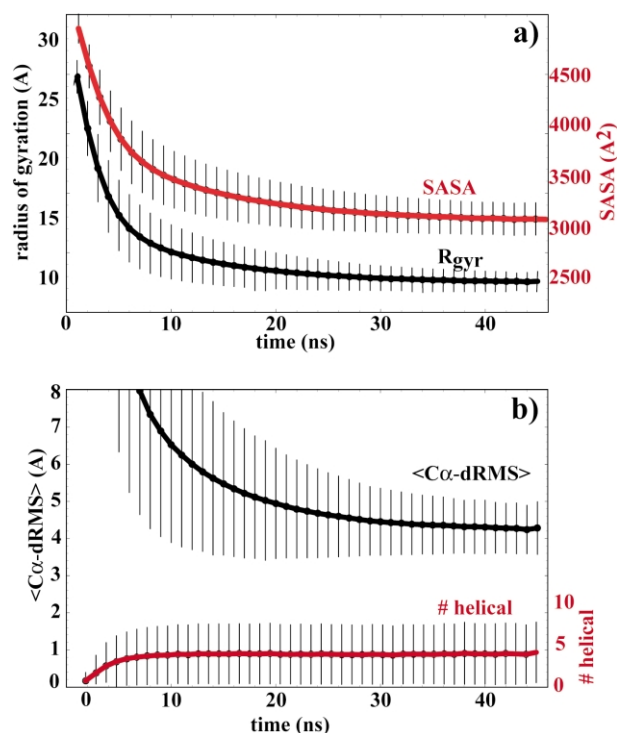


Figure 3. Summary of the folding simulations started from the extended configuration. (a) Radius of gyration (R_{gyr}) and SASA over time for the folding simulations. (b) $\langle C^\alpha$ dRMS) from the mean native matrix at 20 ns (see the text), black, and the total number of helical residues as determined by DSSP,⁴⁰ red, over time. The graphs are based on all of the trajectories that have reached a given time-point: only a small fraction (see the text) of these trajectories have actually resulted in complete folding. This Figure shows the properties of the folding ensemble tens of nanoseconds after initiation of folding.

complete collapse of the unfolded ensemble to the radius of gyration and the SASA of the native state within 20 ns (compare with Figure 2). Due to the stochastic nature of the folding process, in this large ensemble of trajectories exploring essentially the unfolded state, there was a small, but significant number of trajectories (i.e. 35, see below) that have reached the folded state. Figure 4 details the nature of one successful folding trajectory[†]. We start from a completely elongated structure, and quickly (after ~ 5 ns) see relaxation into a random-walk unfolded state. This unfolded structure further collapses monotonously until it forms a compact globular structure with the same radius of gyration and SASA of the native state (~ 17 ns). As mentioned above, most of our simulations end in this collapsed intermediate state in about 20 ns (Figure 3). Since this time is much shorter than the experimentally measured folding time (~ 10 μ s), our simulations suggest that the

[†] For a movie of this trajectory please visit: <http://folding.stanford.edu/villin>

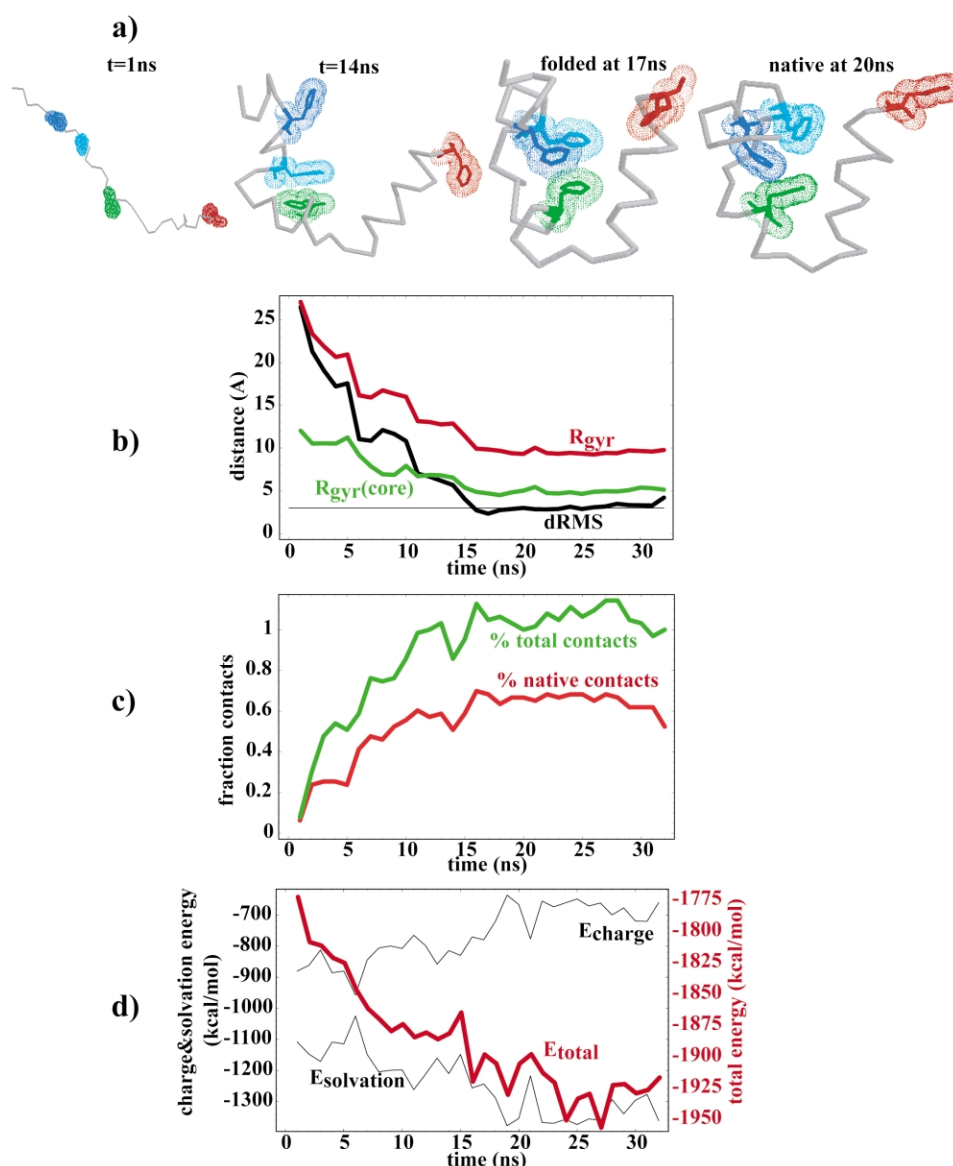


Figure 4. Anatomy of a successful folding trajectory. (a) The first three frames capture the configuration of the villin chain as it folds over time. The backbone is in gray, the phenylalanine residues are in color. The fourth frame captures the representative folded structure from the equilibrium simulations (see Figure 2). The C^{α} dRMS = 2.3 Å and the main-chain RMSD = 3.8 Å between the two structures. (b) Geometric parameters along the trajectory: radius of gyration (R_{gyr}), radius of gyration of the phenylalanine core ($R_{\text{gyr}}(\text{core})$), and C^{α} dRMS from the mean native structure at 20 ns (see the text). (c) The fraction of native contacts (with respect to the experimental 1VII structure¹⁸), red, along the trajectory (red), and the total number of contacts, green, expressed as a fraction of the total number of contacts in the experimental 1VII structure. (d) Total internal energy along the trajectory, red, together with the solvation and the charge–charge components, black.

unfolded ensemble of villin under folding conditions is essentially a compact, hydrophobically collapsed globular structure. This early folding intermediate consists of an ensemble of many conformations with some small amount of native secondary structure, but with incorrect backbone topology and side-chain packing. Unlike the majority of the folding runs, the molecule whose trajectory is shown in Figure 4 folds concomitantly with the hydrophobic collapse. The lowest C^{α} dRMS (~ 2.3 Å) from the folded state is reached around the same time the overall radius of gyra-

tion and the core radius of gyration reach their final, native values (Figure 4(b)).

Among the trajectories that reach low dRMS from the native state, exhibit significant amounts of secondary structure, but have non-native side-chain packing, one particular interaction stands out. Namely, in these runs, Phe36, which in the experimental structure is exposed to the solvent, tends to pack against the rest of the phenylalanine residues (residues number 7, 11 and 18) in the core of the protein. The SASA of Phe36 in the native structure is 224 \AA^2 , compared with

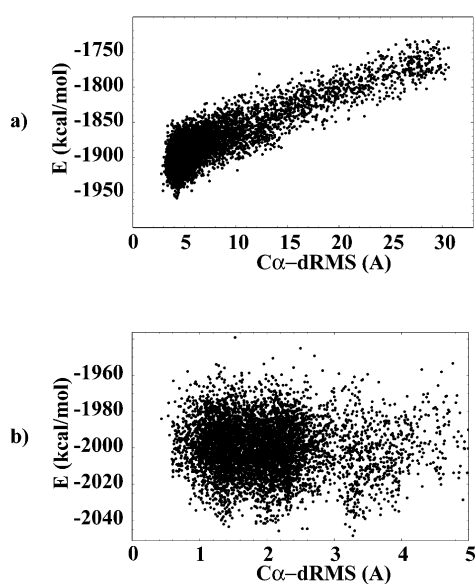


Figure 5. Comparison between the total internal energy and the C^α dRMS from the experimental 1VII structure for (a) the structures in the folding simulations; and (b) the structures below 5 Å in the simulations started from the native state. For clarity, only a randomly chosen 3% of all the structures from the folding simulations, and 8% of all the structures from the native simulations are shown (the key features of the two Figures do not change when all the data are included).

$145(\pm 45) \text{ \AA}^2$ in the ensemble of all the structures in our simulations that are less than 4 Å C^α dRMS away from the native structure (a total of 40,798 structures). In fact, in only about 8% of this compact ensemble from our simulations is Phe36 solvated to the same degree as in the native structure ($\text{SASA} > 200 \text{ \AA}^2$). The interaction between Phe36 and the other phenylalanine residues, due most likely to hydrophobicity and aromatic stacking, needs to break in order for the protein to fold. Our simulations that fold successfully, such as that shown in Figure 4, appear to fold as quickly as they do partly because they, just by chance, steer clear of this potential trap. Since Phe36 forms non-native (misfolded) contacts in our intermediate state (as well as the intermediate found in the Kollman simulation⁶), we predict that removing this bulky hydrophobic side-chain would potentially increase the folding rate.

The successful folder shown in Figure 4 forms about 70% of native contacts (Figure 4(c), bottom), with approximately the same total number of contacts as found in the native state (Figure 4(c), top). These numbers are characteristic of all the other successful folding trajectories. The total energy (Figure 4(d)) of the molecule decreases monotonously, with an apparent plateau coinciding with the presence of an early semi-compact intermediate (7–17 ns). However, it is important to mention that we cannot use the total internal energy as an indicator of folding: while energy does discriminate the folded state within trajec-

tories that reach the folded state, there are many other compact, non-native structures from other trajectories that exhibit the same total internal energy as the final folded state in our simulations. Thus, based on energy alone, we cannot tell which structures are folded and which are not. We demonstrate this in Figure 5(a), where we plot the total internal generalized Born/surface area (GB/SA) energy against the dRMS from the experimental structure for all the structures from the folding simulations. While the dRMS is correlated with the total energy over a wide range, the spread of dRMS values for the low-energy structures is so large that identifying the folded molecules based on energy alone is impossible. The same effect is seen for the simulations started from the native state, where the energies of all the structures below 1 Å dRMS cover a range of almost 100 kcal/mol (1 cal = 4.184 J) (Figure 5(b)).

As mentioned above, our simulations started from the extended state have resulted in a small ensemble of folded structures. In Figure 6 we compare this folded ensemble, which consists of 35 trajectories that have come closest to the mean native state in the C^α dRMS sense (each trajectory has reached a structure with C^α dRMS < 2.85 Å†) with the mean native structure. Our analysis includes all of the structures from these 35 trajectories after the 20 ns time-point (the point at which the unfolded ensemble collapses completely) for a total of 601 structures. Some of the molecules in our simulations fold slightly before the 20 ns time-point, and some after it: since it is actually quite difficult and sometimes even arbitrary to pick the exact moment when a given molecule has folded, we found it simplest to pick a fixed cutoff (i.e. the 20 ns time-point) and analyze all of the structures in the 35 most successful trajectories that came after it. The set of 601 structures chosen in this way is what we term the folded ensemble. Figure 6 shows the distributions of individual C^α and C^β dRMS from the mean native structure for this compact ensemble. As can be seen, individual structures exhibit somewhat broad distributions centered around 3.6 Å and 4.1 Å for C^α and C^β dRMS, respectively. However, when we look at the mean folded matrix (based on all 601 folded or semi-folded structures that came after 20 ns in the

† The exact choice of the C^α dRMS cutoff of 2.85 Å defining the folded state was picked so that the $\langle C^\alpha \text{ dRMS} \rangle$ of the entire folded ensemble, 3.6 Å, falls exactly within $\langle C^\alpha \text{ dRMS} \rangle$ plus one standard deviation determined over the entire simulated native ensemble at 20 ns. In the simulated native ensemble at 20 ns, the $\langle C^\alpha \text{ dRMS} \rangle$ with respect to the mean native structure at that time-point is 2.4 Å with a standard deviation of 1.2 Å (2.4 Å + 1.2 Å = 3.6 Å). As discussed in the text in the context of rate determination, this definition of the folded ensemble is somewhat arbitrary. However, most conclusions and interpretations presented in the text (most importantly regarding the mean structure) do not depend on the exact definition of the folded state.

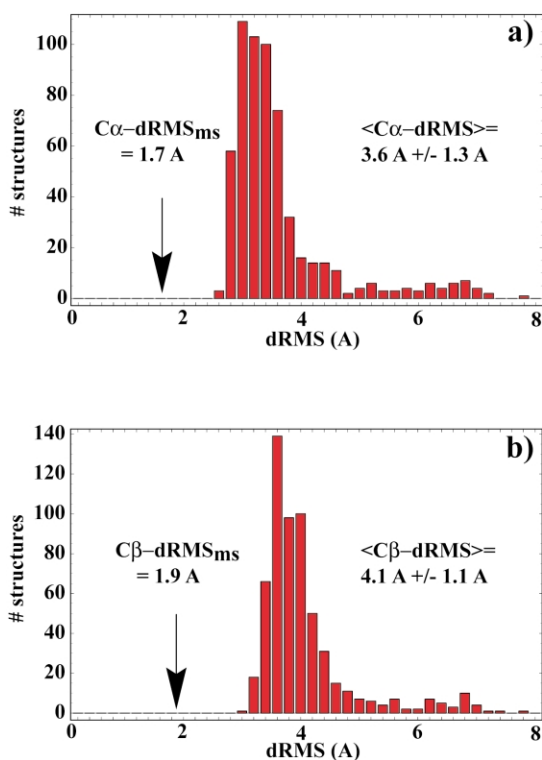


Figure 6. Comparison between the dRMS distributions for the folded ensemble and the dRMS of the mean folded structure based on the same ensemble, all with respect to the mean native structure at 20 ns from the native simulations. The folded ensemble, generated in the folding simulations started from the fully extended state, consists of 601 individual structures from the 35 trajectories that have come closest to the native structure (see the text for details). Here, we compare this folded ensemble with the mean native matrix in two ways. First, we calculate the dRMS from the mean native matrix for each individual member of the folded ensemble: this results in the depicted distributions. Second, we average the distance matrices of all 601 structures in the folded ensemble to get one mean matrix, and then calculate the dRMS between this matrix and the mean native matrix. This results in the values marked by the arrows (dRMS_{ms}, where subscript ms stands for the mean structure). We show the C α dRMS values in (a), and the C β dRMS values in (b). On the right, we show the means and the standard deviations of the depicted distributions.

35 most successful trajectories) and compare it with the mean native matrix, we obtain significantly lower dRMS values (1.7 Å and 1.9 Å for C α and C β , respectively). In other words, the mean folded structure from the simulations is much more similar to the mean native structure than any of the individual folded structures. Since it is the mean structure that would be observed in an ensemble-averaged NMR experiment, we can confidently say that we have indeed folded villin to 1.7 Å and 1.9 Å C α and C β dRMS, respectively.

Finally, we can use our simulations to estimate the folding rate of villin. In any real experiment, the exact value one gets for the folding rate will depend on the observable monitored, sometimes

even to a significant degree (D. Raleigh, personal communication). Similarly, the rate estimate based on our simulations depends critically on the precise definition of the folded state: taking the degree of burial of the hydrophobic core as the relevant order parameter may result in a significantly different rate from, say, using a more geometric parameter such as dRMS. That different properties may show different exponential rates is indicative of downhill, glassy folding; a demonstration of this on a helical system was given recently in a theoretical study by Wolynes and co-workers.¹⁹ To obtain an order of magnitude estimate of the folding rate, we have used C α dRMS with respect to the mean native matrix in defining the folded state: a molecule is folded if C α dRMS < 2.85 Å. As mentioned above, our simulations have resulted in 35 independent simulations (out of approximately 6000 simulations), which satisfy this criterion. On the basis of our data, we can estimate the folding rate and time constant in the following way. We assume, as seen experimentally, that the folding of villin exhibits single exponential behavior. In other words, the probability that a molecule has folded by time t equals:

$$P_{\text{folded}}(t) = 1 - \exp(-kt)$$

where k corresponds to the folding rate. In the limit of $t \ll 1/k$, this expression can be simplified to:

$$P_{\text{folded}}(t) = kt$$

For an ensemble of independent folding processes, the probability of folding, $P_{\text{folded}}(t)$, corresponds simply to $N_{\text{folded}}/N_{\text{total}}$, where N_{total} is the total number of folding processes, and N_{folded} is the number of folding processes that have folded by time t . From this, it follows that the folding rate can be estimated as:

$$k = P_{\text{folded}}(t)/t = N_{\text{folded}}/(N_{\text{total}}t)$$

and:

$$\tau = \text{time constant} = 1/k = (N_{\text{total}}t)/N_{\text{folded}}$$

This means that in ~ 30 ns, which is the average time-span covered by each independent run, we have managed to fold 35/6000 or ~ 0.006 of the entire folding ensemble, resulting in the rate $k = 2 \times 10^5 \text{ s}^{-1}$ or the time constant $\tau \sim 5 \mu\text{s}$. Given the assumptions and simplifications that went into this value, it should be regarded only as an order of magnitude estimate. However, it should be noted that this number is in good agreement with the value obtained by NMR lineshape analysis ($\sim 11 \mu\text{s}$, D. Raleigh, personal communication). As mentioned, the rate estimate from the simulations depends on the exact definition of the folded state. For example, the C α dRMS cutoff of 2.7 Å (satisfied by 11 trajectories) gives the time constant of $\sim 1.5 \mu\text{s}$, while the C α dRMS cutoff of 3 Å (satisfied by 100 trajectories) gives the time constant of $\sim 14 \mu\text{s}$. These results clearly exemplify the difficulties associated with rate estimation from

simulations, and show that any value is accurate to an order of magnitude, at best. The most accurate way of comparing calculated rates with experiment would, of course, involve defining the folded state in the same way as the experimental technique in question (say, by the degree of quenching of a fluorescent probe, by the NMR lineshape width or some other observable). The work is ongoing in our and other laboratories to address this important challenge.

Comparison to protein-folding theory and experiment

The initial collapse of the unfolded ensemble in our simulations appears to be driven by hydrophobic interactions: in approximately 20 ns, the unfolded ensemble compacts to the radius of gyration and the SASA of the native state (Figure 3). Except for a small fraction of fast-folding molecules, which fold to the native state concomitantly with the collapse, most other members of the ensemble collapse into a non-native fold where individual molecules are characterized by incorrect backbone topology, non-native side-chain packing and only nominal amounts of secondary structure. It has been proposed that hydrophobic collapse may play a key role in reducing the size of the conformational space that the molecule needs to explore in order to find the native fold.^{20,21} On the other hand, the compactness of the collapsed globule may severely slow the rate of sampling of the conformational space, so it is not clear *a priori* what the net effect on the folding rate would be. It will be interesting to see if further studies can resolve these two possibilities.

In a significant number of collapsed structures in our unfolded ensemble, we have observed Phe36 packing incorrectly with the rest of the hydrophobic core. We suspect that breaking of this non-native interaction may in fact play a significant role in determining the folding rate of villin. In a preliminary study,²² we have indeed observed a folding event in which this interaction first formed and then broke cooperatively leading to complete folding. In their study of villin folding, Duan & Kollman⁶ have noted the presence of this non-native interaction in their early intermediates, increasing our confidence that it may indeed be physically important. Klein-Seetharaman and collaborators have recently reported an experimental study suggesting that non-native hydrophobic interactions may play a key role in stabilizing early folding intermediates.²³ To avoid aggregation and proteolytic degradation, the nascent peptide collapses quickly, buries the hydrophobics non-discriminately, and then rearranges slowly to the native fold. This proposal fits well with our observation: early packing of Phe36 with the rest of the core may stabilize the collapsed unfolded state, while folding occurs by breaking of this non-native interaction. Plotkin has recently presented a theoretical study of the two conflicting effects of collapse

on folding rate.²⁴ By using the energy landscape arguments, he suggests that the presence of non-native interactions in the collapsed state may actually increase the overall folding rate.

In light of the dominant hydrophobic collapse we observe for villin, it is important to ask if this may be an artifact of the force-field and the solvent model used. The GB/SA solvent model used in the study is based on Still's formulation of the effective Born radii,²⁵ which in turn was parametrized on the basis of thermodynamic data on small molecules. We have not reparametrized the model to include the data for villin, and in this sense the present study is a test of the predictive power of the model. The final verification of the model should be based on comparison with experiment as well as the simulation results obtained with different solvent models, but this is beyond the scope of this study. While it is possible that the fast collapse observed might be an artifact, an encouraging fact is that the estimated folding rate based on our simulations agrees well with experimental data (see above).

It is clear that any potential set employed to model atomic interactions will have its limitations. The relevant question to ask is: how good do they need to be and what would result from errors in these potentials? Since we do see folding to the native state, it appears that the potentials we used were sufficient in this case. Furthermore, our analysis of the mean folded structures suggests that potential sets may even be better than suspected previously. Most of the evaluation of potential sets used in protein simulations so far has been carried out on short single-molecule trajectories. As we have shown, the average properties of ensembles of molecules exhibit significantly different characteristics compared to individual molecules. Since most of the structural information about proteins comes from ensemble-averaged experiments, we believe strongly that simulations should be analyzed on an ensemble level, if fair comparison between experiment and theory is to be attempted. As our results show, the ensemble-averaged results of the simulations are indeed much closer to the desired native data than individual structures and trajectories would lead one to expect. This gives us optimism that potential sets may in fact be accurate enough for detailed structural studies.

It is interesting to speculate why the potential sets would be good enough to simulate structure, and furthermore ask does this tell us anything about the underlying physical reality. Probably the most accurately parameterized parts of the existing protein force-fields are geometry and sterics (bond lengths, bond angles and the hard sphere geometry of atoms). It has been proposed that folding may indeed be determined largely by local steric interactions:^{26,27} if this is the case, then it is easy to understand why accurately parametrized geometry and sterics may be sufficient to bring about correct structures in our simulations. It is

important to emphasize that here we refer to the structural, configurational facets of our computational models. It is quite possible that, while the potential sets are good enough for structural aspects, they distort the dynamic and mechanistic aspects of the folding process. Further work will be needed on both theoretical and experimental fronts in order to address these timely questions in a proper fashion.

The fact that the properly averaged mean structure is closer to the experimental structure than most individual structures comprising the average (Figures 2(c) and 6) has methodological significance, and may actually tell us something important about the nature of protein structure in general. Namely, this result suggests that structures based on ensemble measurements may actually hide a much greater degree of underlying diversity than is usually suspected. In other words, the free energy well belonging to the native state may actually be quite broad and may contain a fairly diverse set of structures: the final, experimentally determined, refined structures may just be idealized averages. This fact has been addressed in the literature under different guises. Van Gunsteren and collaborators have used molecular dynamics (MD)-based time-averaged restraints in structure refinement and showed that an ensemble of structures that satisfies experimental constraints only on average (with many individual structures actually violating some of the constraints) leads to the refinement of higher-quality structures compared to other refinement procedures.^{28–30} Van Gunsteren's group has shown how, in particular, the NMR-derived constraints are actually quite insensitive to the nature of the underlying ensemble.^{31,32} This supports the claim that the average structure of the native state, at least as determined by NMR methods, may conceal a significant degree of diversity characterizing the underlying microscopic ensemble. Shortle's group has gone even a step beyond, in suggesting that even the topology of the denatured state may retain the features of the native state.^{26,33–35} Recently, we have shown how the geometry of the compact unfolded states may on average be extremely native-like for small proteins.¹⁷ Further, Shortle *et al.* have shown in their *ab initio* structure prediction study that clusters with the largest number of similar structures in their decoy sets often contain the native structure.³⁶ The converse of this is that the native basin may actually be quite broad. Finally, Onuchic has suggested that the native state is the center of the folding funnel.³⁷ Our study, we believe, complements the above results and goes a step beyond, in that (a) we propose a way of averaging ensembles of structures by generating mean distance matrices, (b) we generate a folded ensemble by using unbiased simulations, and (c) we indeed show how the mean structure from folding simulations matches the simulated native structure better than any individual structure comprising the mean (Figure 6).

To summarize, using existing force-fields and solvent models, MD simulation, and large-scale distributed computing techniques, we have simulated the folding of a small protein (the villin head-piece) starting from an unfolded state with only the knowledge of the primary sequence of the protein. Our ability to simulate the folding of villin is, in a sense, a validation of potential sets and our distributed computing methodology. We have shown that while clearly having limitations, current potential sets may be accurate enough to see folding in at least small proteins. However, since the energies of our model do not provide a distinguishable native basin (Figure 5), we were forced to rely on the knowledge of the native structure to identify the folded structures in our simulations. Improvement of the biomolecular force-fields to allow identification of the native structures based on energy alone is an area where further improvements could be made, and we hope distributed computing paradigms will play an important role in this. However, since the native state is a minimum in free energy and not potential energy, it is possible that even a perfectly accurate, physically realistic potential energy function could not be used to identify the folded state, since the entropic component of the free energy may prove to be a critical component. Also, one would expect the fractional energy fluctuations to be of the order of the inverse square-root of the number of atoms (which is quite significant for small systems, such as proteins). The unprecedented degree of sampling provided by distributed computing may help address this issue as well. Furthermore, we have shown that ensemble averaging of simulation results may be very important for proper comparison between theory and experiment. Finally, our computational method appears to be a viable means to simulate biomolecular dynamics on time-scales tens of thousands of times longer than that available to a fast workstation and ten to a hundred times longer than a typical super-computer. This computational advance will greatly facilitate a detailed study of multiple atomistic folding trajectories in order to discern the folding mechanism with statistical certainty.

Methods

Using a heterogeneous computer cluster we have generated thousands (see above) of short (tens of nanoseconds) independent trajectories for the 36-residue villin headpiece segment. 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 structure of the molecule (villin, PDB code 1VII,¹⁸ sequence: MLSDEDFKAVFGMTRSAFANLPLWKQQ-NLKKEKGLF). Even though the simulations are all started from the same structure (fully extended or native), they quickly diverge from each other due to the stochastic aspects of the Langevin dynamics (see below). We have numbered the residues starting from 1

(Met) to 36 (Phe). Residues 2–36 correspond to residues 791–825 of intact chicken villin and to residues 42–76 of the C-terminal 76 amino acid residue chicken villin headpiece domain. Met1 residue is the N-terminal residue from the expression system and is not from villin. However, it was included in the 1VII experimental structure, and we include it here. The simulations, run using the Tinker biomolecular simulation package†, involved Langevin dynamics in implicit GB/SA²⁵ solvent with a 2 fs integration step at 300 K with the water-like viscosity of $\gamma = 91 \text{ ps}^{-1}$. Bond lengths were constrained using RATTLE.³⁸ No cutoff was used for electrostatics. The protein was modeled using the OPLSua force-field.³⁹ Coordinates were output every 1 ns. To compare structures (i.e. distance matrices) we have used dRMS, defined as:

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

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. Contacts (Figure 4) were defined in the following way: two residues are in contact if they are at least three or more sequence positions apart and if their C^α atoms are within 7 Å of each other. The simulations were carried out on 10,000+ processors as a part of our ongoing Folding@Home distributed computing project‡, and involved a total of about a quarter of a trillion (2.5×10^{11}) integration steps. This corresponds to approximately 1000 single CPU (500 MHz) years of computation.

Acknowledgements

We especially thank the thousands of Folding@Home contributors, without whom this work would not be possible. A complete list of contributors can be found at <http://folding.stanford.edu>. We thank Robert Baldwin for useful comments. We thank Dan Raleigh and collaborators for their unpublished results on the experimental folding time for this 36-residue villin fragment, and Jay Ponder for the use of and help with the Tinker MD code. B.Z. and C.S. each acknowledge support from an HHMI predoctoral fellowship. M.S. acknowledges support from a Hertz Foundation and a Stanford Graduate predoctoral fellowships. This work was supported by grants from the ACS PRF (36028-AC4), 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.

† <http://dasher.wustl.edu/tinker/>

‡ <http://folding.stanford.edu>

References

- Dobson, C. M., Sali, A. & Karplus, M. (1998). Protein folding: a perspective from theory and experiment. *Angew. Chem. Int. Ed. Engl.* **37**, 868–893.
- Dill, K. A. & Chan, H. S. (1997). From Levinthal to pathways to funnels. *Nature Struct. Biol.* **4**, 10–19.
- Brooks, C. L., III, Gruebele, M., Onuchic, J. N. & Wolynes, P. G. (1998). Chemical physics of protein folding. *Proc. Natl Acad. Sci. USA*, **95**, 11037–11038.
- Prusiner, S. B. (1998). Prions. *Proc. Natl Acad. Sci. USA*, **95**, 13363–13383.
- Nelson, J. C., Saven, J. G., Moore, J. S. & Wolynes, P. G. (1997). Solvophobic driven folding of non-biological oligomers. *Science*, **277**, 1793–1796.
- 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.
- Shea, J. E. & Brooks, C. L., III (2001). From folding theories to folding proteins: a review and assessment of simulation studies of protein folding and unfolding. *Annu. Rev. Phys. Chem.* **52**, 499–535.
- Tobias, D. J. & Brooks, C. L., III (1987). Calculation of free-energy surfaces using the methods of thermodynamic perturbation-theory. *Chem. Phys. Letters*, **142**, 472–476.
- 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.
- Fersht, A. R. & Daggett, V. (2002). Protein folding and unfolding at atomic resolution. *Cell*, **108**, 573–582.
- 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.
- Wong, K. B., Clarke, J., Bond, C. J., Neira, J. L., Freund, S. M., Fersht, A. R. & Daggett, V. (2000). Towards a complete description of the structural and dynamic properties of the denatured state of barnase and the role of residual structure in folding. *J. Mol. Biol.* **296**, 1257–1282.
- 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.
- Shirts, M. R. & Pande, V. S. (2001). Screensavers of the world, unite!. *Science*, **290**, 1903–1904.
- Shirts, M. R. & Pande, V. S. (2001). Mathematical analysis of coupled parallel simulations. *Phys. Rev. Letters*, **86**, 4983–4987.
- 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.
- Zagrovic, B., Snow, C. D., Khaliq, S., Shirts, M. R. & Pande, V. S. (2002). Native-like mean structure in the unfolded ensemble of small proteins. *J. Mol. Biol.* **323**, 153–164.
- 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.
- Hardin, C., Luthey-Schulten, Z. & Wolynes, P. G. (1999). Backbone dynamics, fast folding, and secondary structure formation in helical proteins

- and peptides. *Proteins: Struct. Funct. Genet.* **34**, 281–294.
20. Ptitsyn, O. B. (1995). Molten globule and protein folding. *Advan. Protein Chem.* **47**, 83–229.
 21. Fersht, A. R. (1999). *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*, Freeman, New York.
 22. Pande, V. S., Baker, I., Chapman, J., Elmer, S., Khaliq, S., Larson, S. *et al.* (2002). Atomistic protein folding simulations on the hundreds of microsecond time-scale using worldwide distributed computing. *Biopolymers*. In press.
 23. Klein-Seetharaman, J., Oikawa, M., Grimshaw, S. B., Wirmer, J., Duchardt, E., Ueda, T. *et al.* (2002). Long-range interactions within a nonnative protein. *Science*, **295**, 1719–1722.
 24. Plotkin, S. S. (2001). Speeding protein folding beyond the Go model: how a little frustration sometimes helps. *Proteins: Struct. Funct. Genet.* **45**, 337–345.
 25. 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.* **101**, 3005–3014.
 26. Shortle, D. & Ackerman, M. S. (2001). Persistence of native-like topology in a denatured protein in 8 M urea. *Science*, **293**, 487–489.
 27. Srinivasan, R. & Rose, G. D. (1999). A physical basis for protein secondary structure. *Proc. Natl Acad. Sci. USA*, **96**, 14258–14263.
 28. van Gunsteren, W. F., Brunne, R. M., Gros, P., van Schaik, R. C., Schiffer, C. A. & Torda, A. E. (1994). Accounting for molecular mobility in structure determination based on nuclear magnetic resonance spectroscopic and X-ray diffraction data. *Methods Enzymol.* **239**, 619–654.
 29. Gros, P., van Gunsteren, W. F. & Hol, W. G. (1990). Inclusion of thermal motion in crystallographic structures by restrained molecular dynamics. *Science*, **249**, 1149–1152.
 30. Torda, A. E., Brunne, R. M., Huber, T., Kessler, H. & van Gunsteren, W. F. (1993). Structure refinement using time-averaged *J*-coupling constant restraints. *J. Biomol. NMR*, **3**, 55–66.
 31. 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.
 32. 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.
 33. 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.
 34. 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.
 35. 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.
 36. 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.
 37. Onuchic, J. N. (1997). Contacting the protein folding funnel with NMR. *Proc. Natl Acad. Sci. USA*, **94**, 7129–7131.
 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 B. Honig

(Received 15 April 2002; received in revised form 6 September 2002; accepted 10 September 2002)