

β -Hairpin Folding Simulations in Atomistic Detail Using an Implicit Solvent Model

Bojan Zagrovic¹, Eric J. Sorin² and Vijay Pande^{1,2*}

¹*Biophysics Program*

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

We have used distributed computing techniques and a supercluster of thousands of computer processors to study folding of the C-terminal β -hairpin from protein G in atomistic detail using the GB/SA implicit solvent model at 300 K. We have simulated a total of nearly 38 μ s of folding time and obtained eight complete and independent folding trajectories. Starting from an extended state, we observe relaxation to an unfolded state characterized by non-specific, temporary hydrogen bonding. This is followed by the appearance of interactions between hydrophobic residues that stabilize a bent intermediate. Final formation of the complete hydrophobic core occurs cooperatively at the same time that the final hydrogen bonding pattern appears. The folded hairpin structures we observe all contain a closely packed hydrophobic core and proper β -sheet backbone dihedral angles, but they differ in backbone hydrogen bonding pattern. We show that this is consistent with the existing experimental data on the hairpin alone in solution. Our analysis also reveals short-lived semi-helical intermediates which define a thermodynamic trap. Our results are consistent with a three-state mechanism with a single rate-limiting step in which a varying final hydrogen bond pattern is apparent, and semi-helical off-pathway intermediates may appear early in the folding process. We include details of the ensemble dynamics methodology and a discussion of our achievements using this new computational device for studying dynamics at the atomic level.

© 2001 Academic Press

Keywords: protein folding; simulation; ensemble dynamics; β -hairpin; folding mechanism

*Corresponding author

Introduction

Understanding protein folding is one of the much desired goals of modern day molecular biology.^{1–3} The ability to predict the folding mechanism and the final structure of a protein based on its sequence would dramatically affect different fields ranging from biochemistry to molecular

medicine to nanotechnology. After decades of largely independent experimental and theoretical work, the field of protein folding is slowly but undeniably entering a mature age in which the two are converging. Experimental techniques have become sophisticated enough to probe the folding of small, fast-folding proteins and protein elements, while computational power and algorithms have reached a level at which simulating these events is tractable.

Recently, folding of the C-terminal β -hairpin from protein G (Figure 1) has received much attention from both experimental and theoretical fronts.^{4–19} Parallel and antiparallel β structures are, together with α helices, the key secondary structural elements in proteins, and it is believed that understanding the folding of these elements will be a foundation for investigating larger and more complex structures. The study of isolated β -sheets has for a long time been limited by the lack of an amenable experimental system. The breakthrough experiments by the Serrano^{5,6} and

B. Z. and E. J. S. contributed equally to this work.

Abbreviations used: d_{\min} , minimum distance between hydrophobic core residues; F , relative statistical energy; PC, principal component; MC, Monte Carlo; NOE, nuclear Overhauser enhancement; MD, molecular dynamics; Ace, acetyl; ASA, solvent-accessible surface area; E, fully extended conformation; H, hydrophobically collapsed intermediate; F, folded conformation; U, more collapsed unfolded conformation; GB, generalized Born model; SA, surface area; PB, Poisson-Boltzmann.

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

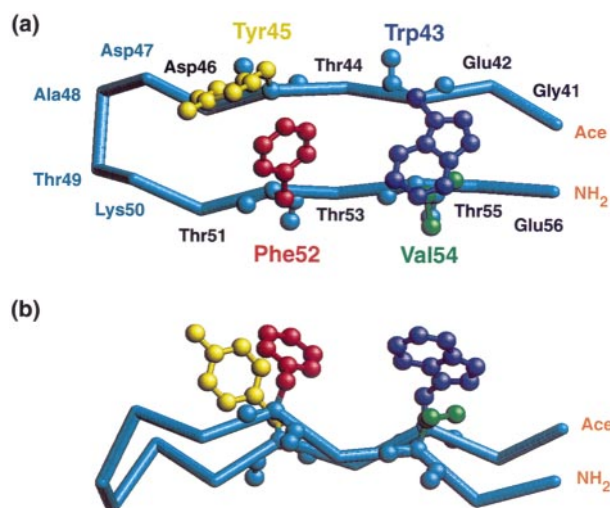


Figure 1. β -Hairpin structure from the 1GB1 NMR structure of protein G.⁴ In (a) the “top” view of the hairpin structure shows the relative strand positions and the hydrophobic side-chains which comprise the core: Trp43 (blue), Tyr45 (yellow), Phe52 (red), Val54 (green). Note that the numbering system agrees with the original nomenclature of the protein G NMR structure.⁴ The excised hairpin is expected to exhibit slightly different properties, as tertiary interactions with the rest of the protein G molecule are absent when the hairpin is alone in solution. (b) The same representation rotated $\sim 90^\circ$ about the long axis of the hairpin. Numerous characteristics of the structure are clearly represented: hydrophobic packing occurs on only one side of the hairpin and is located near the central region of the strands; β -sheet angles are present along most of the chain, with the loop and terminal residues breaking this trend; a slight twist about the channel axis is evident.

Eaton⁷ groups have recently established the β -hairpin from the C terminus of protein G as the system of choice to study β -structures in isolation. The kinetic studies by the Eaton group⁷ have revealed several surprising facts, the most important being the observation that the hairpin folds in a manner that is very similar to the folding of small proteins: in their thermal unfolding experiments, the peptide behaved as a two-state system, both kinetically and thermodynamically. Together with the fact that the hairpin possesses a hydrophobic core which is well packed in the folded state, these results have added an additional impetus for studying the folding of this molecule computationally.

Based on the high-resolution structures of protein G^{4,20–22} and the NMR analysis of the hairpin alone in solution,⁵ the structural features of the hairpin include two β -pleated strands joined by a six residue turn, a closely packed hydrophobic core (residues 43, 45, 52 and 54), and a number of inter-strand hydrogen bonds connecting the two opposing strands of the hairpin. It is important to note that as of now there is no high-resolution experimental structure of the hairpin alone in solution: the number of nuclear Overhauser enhancements (NOEs) from the NMR analysis by the Serrano group was not sufficient to fully constrain the structure.⁵ Figure 1 shows the fold of the hairpin in protein G (PDB ID code 1GB1, first model), as determined by NMR.⁴ In the absence of a high-resolution structure of the hairpin alone, this best serves as an approximate model for the interactions and structural characteristics that would be

expected from an excised hairpin in solution. One should keep in mind, however, that the measurements by the Serrano group do indicate that, despite the overall similarity, the structure of the hairpin alone in solution may differ from the high-resolution structures of the hairpin embedded in protein G.⁵ In Figure 1(a) the hairpin backbone is displayed in cyan, with the four hydrophobic side-chains colored to show their close packing (Trp43 in blue, Tyr45 in yellow, Phe52 in red and Val54 in green). This same structure has been rotated $\sim 90^\circ$ about the hairpin axis in (b) to allow easy visualization of the C^α trace. β -Sheet angles and a slight twist about the channel axis are evident. This side view also demonstrates the hydrophobic packing that occurs only on one side of the hairpin, near the center of the folded structure.

Being the smallest known naturally occurring system which exhibits many features of a full size protein, the hairpin has recently motivated a number of theoretical studies involving an impressive array of different computational techniques.^{8–12,15–18} Nevertheless, simulating the folding of the molecule in atomistic detail at a physiological temperature has not been previously achieved. The hairpin folds with a time constant of several microseconds,⁷ almost three orders of magnitude longer than the typical time-scales attainable by molecular dynamics simulations. Due to this significant time-scale gap, there are a number of important questions regarding the folding of the hairpin which are still controversial.¹⁹ The foremost among these concern (a) the nature of the on-pathway intermediate observed in some, but not all,

studies of this peptide; (b) the existence of off-pathway intermediates that may act to hinder hairpin formation; (c) the relative importance of interstrand hydrogen bonds in comparison to hydrophobic core formation; (d) the order of structure formation in the course of folding; and (e) the heterogeneity of the folded state.

Recently, we have introduced a new technique, ensemble dynamics (see Methods), aimed at bridging the gap between computationally achievable time-scales and the much longer time-scales involved in protein folding.^{23–25} Using a supercluster of processors around the world and distributed computing techniques²⁵ we have folded α -helices, the 36 residue villin headpiece and a 20 residue zinc finger molecule (I. Baker *et al.*, unpublished results). Here, we have used the ensemble dynamics method to analyze the folding process of the C-terminal β -hairpin from protein G. We have obtained an ensemble of complete folding trajectories in implicit solvent allowing us to analyze the folding mechanism in detail. In addition, the large amount of data has allowed us to investigate the nature of both the hydrophobic on-pathway and semi-helical off-pathway intermediates that are witnessed in the course of folding.

Results

Before presenting and interpreting our findings, a discussion of the overall character of the data obtained is necessary. Given the nature of the ensemble dynamics method (see Methods;^{23–25}), it is clear that while numerous simulations resulted in completely folded hairpins (a total of 40, with eight of these being fully independent), the majority of Trial simulations resulted only in unfolded or partially folded structures. Naturally, this gives us two data sets to examine. The first is a very large data set containing all of the simulated trajectories and including millions of integration time steps, and will be referred to herein as the composite data set. This ensemble includes few folding events, and represents an unfolded ensemble approximately 14 ns into folding. From this, we can investigate conformational space quite well, having nearly 38 μ s of simulation time. Since this ensemble has not crossed the free energy barriers between the unfolded and folded states a sufficient number of times, and therefore is not an equilibrium ensemble, we could not use it to quantitatively describe the underlying free energy landscape. Nonetheless, we could still use it to determine the approximate locations of the free energy barriers, witness any off-pathway intermediates which may be present, and categorize the structures which comprise it. The second data set to examine is a small but important subset of the composite data. This set consists only of those folding trajectories which resulted in fully formed hairpins and it allows us to examine the kinetics and mechanism of hairpin folding in detail. It is

important to mention that the trajectories which are included in this set are fully independent from one another. Regarding the thermodynamic results we present, consideration must be given to the extent of our sampling. While we have sampled a sizeable fraction of the configurational space, we have by no means fully covered it: after all, our ensemble of generated structures is not nearly in equilibrium. For this reason, we employ the term statistical energy herein to describe the calculated energy surfaces which, in the limit of sufficient sampling, should approach the true free energy landscape of this system. The statistical energy is calculated as the logarithm of the probability of finding a conformation with specific values of the observed folding parameters.

Overview of structural trends and sampling of configurational space

We start with an overview of the major structural trends characterizing the large ensemble within the composite data set. The configurations observed typically fall into one of three families of structures. Folded structures display backbone β -sheet angles, strand symmetry, and a centralized hydrophobic core on one side of the hairpin as seen in the structure of the hairpin from protein G (Figure 1). These structures will be analyzed in more detail below. The unfolded, solvated structures which comprise most of our data show tremendous variation, but most often involve turns and bends which are shifted away from the center of the sequence, with no particular shift being of noteworthy prominence. The semi-helical intermediates witnessed in this study follow suit with small helical turn regions forming in one of many possible ways. Short helical stretches at the C terminus, at the N terminus, at or near the center of the peptide, and simultaneously at each end of a single structure were all observed. A very small number of conformations with mixed (α/β) secondary structure were also seen. Visual analysis of these structures suggests that they are merely random events which occur during the search for low energy among the myriad of conformational possibilities: the β -like portion typically involves only two well paired residues near an end of the sequence, while the semi-helical region most often involves only a single turn. Though we have recorded an event in which a semi-helical intermediate precedes formation of the folded hairpin, we have observed nothing to suggest that the mixed (α/β) structures lie along any given pathway between the semi-helical and folded hairpin regions of conformational space. Finally, these structures were seen very rarely (in less than 0.05% of the generated configurations), and typically persisted for only a few hundred picoseconds. Upon examination of the semi-helical structures within our composite data, the following characteristics of their appearance both support our classification above and clarify our model of their role in

hairpin formation: (a) helical-turn formation appears to be independent of residue identity in this fragment; (b) all structures with substantial helical content ($N_\alpha > 5$) were devoid of sheet-like character; and (c) no mixed (α/β) states were seen in configurations with high hairpin character ($N_\beta > 5$). In the section following this one we will describe a transition into the semi-helical region of configurational space as well as a transition out of it. These events were observed early in one of the successful folding trajectories described below.

Figure 2 plots the statistical energy as a function of the molecule's radius of gyration and RMSD from the 1GB1 structure. When examining the composite data set shown in Figure 2(a), we observe a single energetic descent toward the folded hairpin. The hairpin configuration is noted to have a radius of gyration of less than 7 Å. An analogous statistical energy profile based solely on the fully folded trajectories is plotted in Figure 2(b). Three distinct regions belonging to the fully extended conformation (E), a more collapsed unfolded conformation (U) and the folded conformation (F) are readily identifiable. The hydrophobically collapsed intermediate, H, (see below) has also been labeled in this plot. Note that with respect to the radius of gyration and RMSD degrees of freedom, the H state partially overlaps with both the U and the F states. Examining other degrees of freedom, however, can allow for explicit detection of this intermediate, as demonstrated in Figure 3. Since the unfolded conformation in solution is probably better represented by the partially collapsed ensemble than by the fully extended ensemble, we believe that the experimentally detectable folding transition is the one between states U and H/F in our model. For comparison, we have also calculated an approximate statistical energy profile for the equilibrated, fully folded hairpin from the 1GB1 structure of protein G. This profile, shown in Figure 2(c), is based on a single 15 ns run starting from the 1GB1 structure of the hairpin under the same simulation conditions as the ones used in the ensemble dynamics simulations.

The relative statistical energy (F) as a function of the number of residues with β -sheet backbone dihedral angles and the minimum distance between the hydrophobic core residues for the fully folded trajectories, $F(N_\beta, d_{\min})$, is shown in Figure 3. We require a minimum sheet-like content of $N_\beta = 4$ for fully folded structures (F) and this is observed only for conformations with well packed hydrophobic cores ($d_{\min} < 5$ Å). No folded hairpin structures are observed to have a core separation above this threshold. Furthermore, a large population of structures that have already undergone initiation of core formation (low d_{\min}) but have little β -sheet character are observed. Similarly, there is a small population of structures with few β -sheet residues ($N_\beta \sim 2$) and very little core packing (large d_{\min}). In such unfolded structures, any sheet-like residues are improperly located and

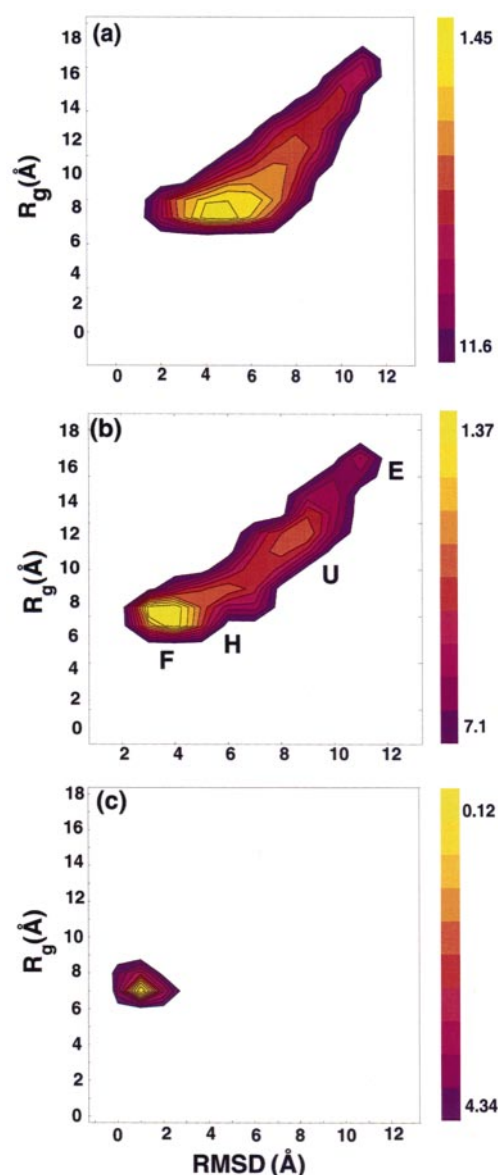


Figure 2. Statistical energy analysis of the folding process. The statistical energy, calculated as the negative logarithm of the fraction of the total population, is shown as a function of the overall RMS distance from the 1GB1 structure (residues 43-54; two residues at each end of the peptide were excluded, since they are expected to be frayed in solution) and the radius of gyration of the molecule for (a) the composite data set, (b) the eight complete and independent trajectories which make up the folded data set, and (c) a 15 ns equilibrium run of the 1GB1 hairpin structure. The composite data set closely mimics the statistical energy surface reported by Dinner *et al.*,¹⁰ showing a smooth descent from the unfolded region into the folded state. Note the stark contrast displayed by the folded ensemble, in which several distinct thermodynamic states are visible: the fully extended state (E), the slightly more compact unfolded state (U), and the native hairpin (F) are all labeled. Note that the H state, which is also labeled here, overlaps with the U state and the F state. The same profile based on a single equilibrium run lasting 15 ns, started from the 1GB1 structure of the hairpin, compares well with the F minimum in (b). In all three graphs, contours are drawn at linearly increasing intervals between the lowest and the highest values shown with each scale bar.

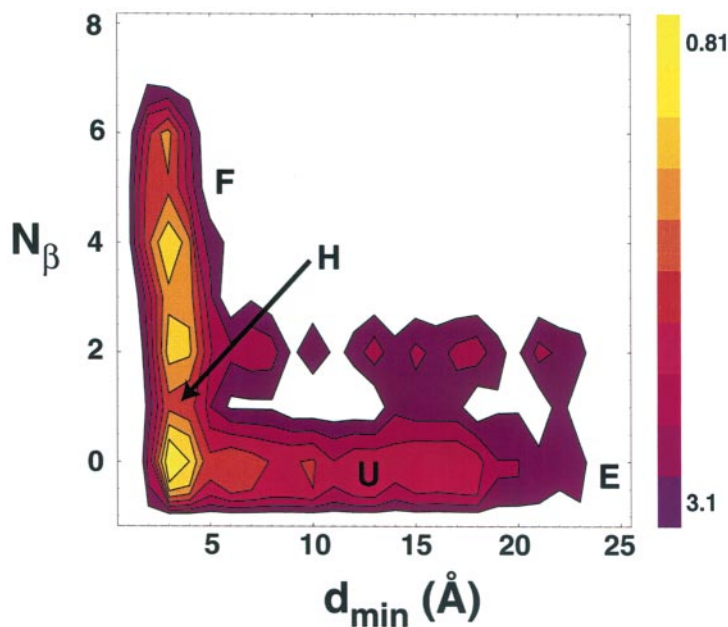


Figure 3. Statistical energy as a function of N_β and d_{\min} . The plot includes only the eight trajectories which resulted in fully formed hairpins. Note the distinct absence of structures that are highly sheet-like yet display large core separations, which demonstrates the vital importance of early hydrophobic interaction, the essential driving force initiating the folding process. The primary minima located at d_{\min} less than 5 Å and N_β from 0 to 2 define the hydrophobic kinetic intermediate (H) which is characterized by low core residue separation (as indicated by the arrow) and lack of hairpin structure. The unfolded (U), extended (E), and folded (F) regions of the surface are labeled as well. The contours are drawn at linearly increasing intervals between the lowest and the highest values shown with the scale bar.

must rearrange to allow for the proper folding geometry to occur. Finally, we note the easily identified minima at low core separations and low β -sheet content. These minima represent the configurations that have reached the primary turning point in the folding process: the hydrophobic contacts are established while proper β -sheet backbone (ϕ, ψ) angles and finalized hydrogen bonds are still not formed. We define this state (discussed below) as a hydrophobic intermediate (H), in agreement with several other computational studies of the hairpin.^{9,10,15}

Kinetics and mechanism

Our simulations resulted in eight independently obtained trajectories of fully folded final structures which exhibit all of the canonical characteristics of the β -hairpin (a well-packed hydrophobic core, a number of interstrand backbone to backbone hydrogen bonds, and β -sheet backbone dihedral angles). All eight structures came from independent Series (consisting of 100 Trial simulations each), meaning that at no point were the Series in any way coupled to each other (see Methods).

A detailed analysis of the folding trajectory for one member of the folded ensemble (Series 2) is shown in Figure 4. The folding process begins with a rapid collapse of the extended peptide to a more globular conformation. This step is characterized by the appearance of one to three temporary

hydrogen bonds as well as the early appearance of a hydrophobic interaction (Trp43-Phe52) which will be key for further formation of the hairpin. During this early collapse, the radius of gyration of the molecule drops essentially to its final value, while the RMSD from the 1GB1 structure reaches a plateau which persists for about 30 generations[†]. During this time, different hydrogen bonds appear and then break and this search for the final structure revolves essentially around the constant hydrophobic interaction mentioned above. At generation 41, one sees a cooperative formation of the fully formed hydrophobic core and of the key hydrogen bonds (Tyr45-Phe52, Trp43-Val54). This event marks the completion of the folding process of the peptide and is accompanied by the RMSD, the total energy, and the total average number of hydrogen bonds all reaching their final, equilibrium values. After the peptide has folded, the hairpin structure remains fairly stable: the total number of backbone-backbone hydrogen bonds fluctuates between three and five (the two key hydrogen bonds mentioned earlier rarely break), the core remains well packed, and the total potential energy hovers around -900 kcal/mol (1 cal = 4.184 J). At generation 155 one can see a minor reorganization of the hydrophobic core which is accompanied by the breaking of the Tyr45-Phe52 hydrogen bond, an increase in the solvation energy and a decrease in the charge-charge energy. Nonetheless, after about ten generations, the fully folded hairpin structure reappears. Figure 4(e) summarizes the main events along the folding trajectory: each time-series of observables was scaled between its highest (yellow) and lowest (violet) values ($X_{\text{final}} = (X - X_{\min}) / (X_{\max} - X_{\min})$),

[†] Time can most naturally be approximated in the following way: each generation corresponds to 100 processors \times 0.1 ns/generation per processor = 10 ns/generation (see Methods for more details).

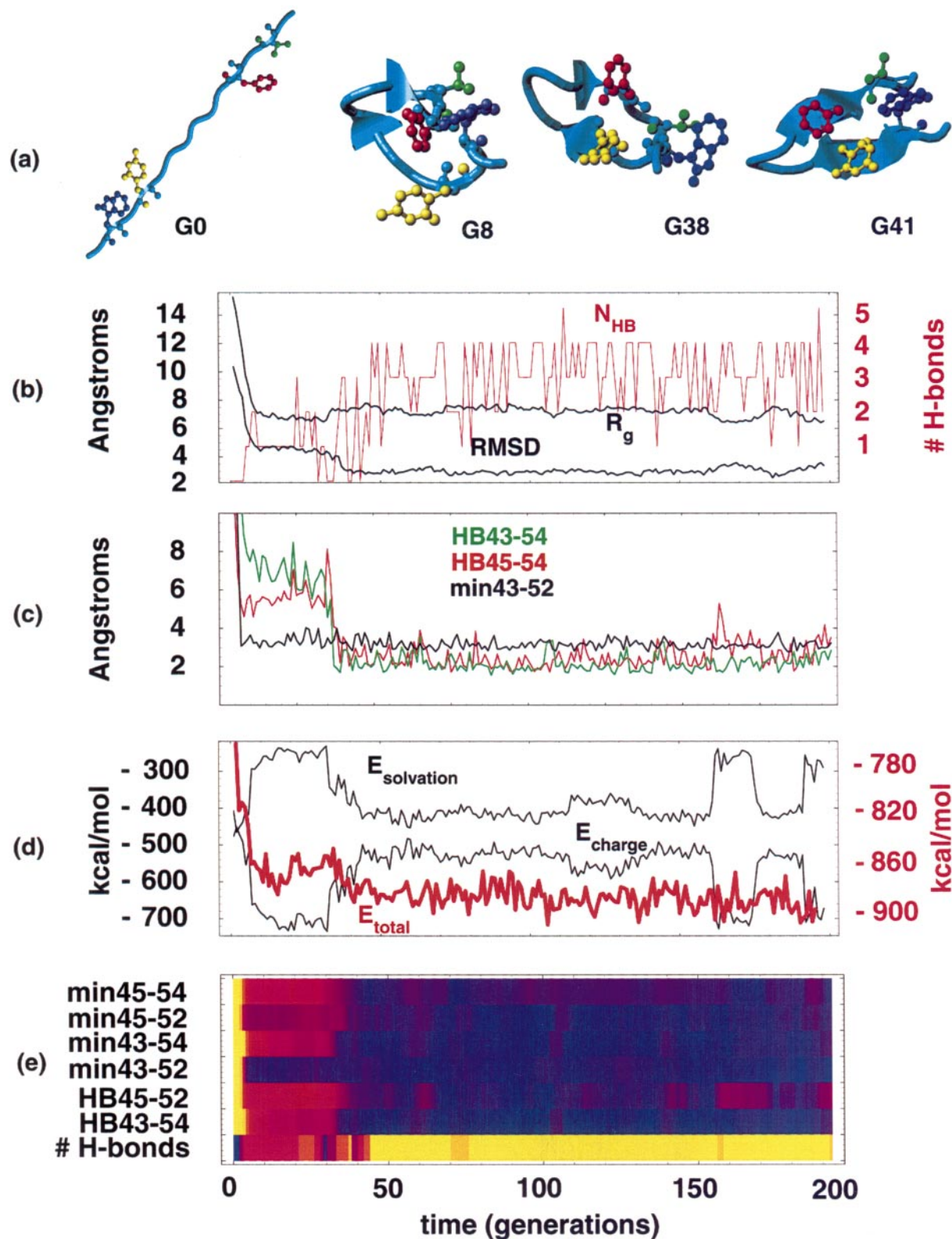


Figure 4 (legend opposite)

where X stands for each observable that was analysed. As is apparent, the minimum distance between the side-chains of Trp43 and Phe52 reaches its low equilibrium value before final

hydrogen bonds are established and before other hydrophobic interactions reach their peaks.

Figure 5 shows a similar plot for each of the eight members of our folded ensemble. For each

simulation, we plot the total number of hydrogen bonds, the minimum distances between the hydrophobic core residues, and the distances between the key hydrogen bonding partners (only those hydrogen bonds which rarely break after the folding is complete are included). All or most of the runs are characterized by a fluctuating total number of hydrogen bonds, and an early appearance of hydrophobic interactions followed by the formation of hydrogen bonds connecting the two strands of the hairpin. As was mentioned before, most hairpin structures differ in the pattern of key interstrand hydrogen bonds even though their hydrophobic cores are packed in a similar fashion. Finally, most successful trajectories exhibit partial unfolding and subsequent refolding after the initial folding event, which we expect properly emulates a real hairpin in solution.

The first two principal components (PCs) (refer to Methods for an analytical description of these degrees of freedom) are plotted *versus* time (in generations) for these same eight trajectories in Figure 6. While $PC_1(t)$ reflects the events of Figure 5 quite well, it is evident from the lack of significant transitions along $PC_2(t)$ that this principal component is less suitable for describing the kinetics of interest. The majority of these eight trajectories, however, show a well defined transition very early in the PC_1 pathway. Note that PC_1 has been standardized to zero for the folded structures. Also note that regarding the formation of β -type backbone angles (arrows above trajectories in Figure 5), even this degree of freedom does not appear capable of distinguishing between the properly folded hairpin and H intermediates which may look hairpin-like but lack the proper dihedral backbone angles. We see that for this study a single folding coordinate, $PC_1(t)$, can be used to represent an entire set of geometric parameters throughout the folding process and gives details of the collapse and core formation, but cannot be trusted to adequately demonstrate that full folding has occurred.

Inset in each PC plot for these eight trajectories are the time-dependent, solvent-accessible surface areas (ASA) of the hydrophobic core. These traces (green) explicitly detail the hydrophobic collapse

observed in all of our folding simulations. Initially, the core residues of the extended structure have a combined solvent-exposed surface area of nearly 700 \AA^2 . In over half of our eight folding trajectories (Series 2, 5, 7, 9, and 17) this area falls to between 400 \AA^2 and 500 \AA^2 very rapidly (within the first 40 generations). Correlating well with other degrees of freedom as shown in both Figures 5 and 6, Series 9 remains folded for only a short period before the core becomes much more solvent-exposed.

Series 11 offers a direct glimpse into the formation of a semi-helical off-pathway intermediate and its disappearance, which ultimately leads to hairpin formation. The first ~ 20 generations shown in Figure 6 for this Series exhibit a temporarily stabilized, high-valued PC_1 alongside a consistently high core solvent-accessible surface area. Figure 7 describes the process of helical-turn formation and relaxation during this period along three degrees of freedom: the total potential energy of the structure, the number of residues meeting the α -helical criterion (N_α), and the number of hydrogen bonds (N_{HB}). Above these plots are structures representing the configurations at certain time points in the simulation. Beginning with an unfolded structure at the second generation (structure G2) we observe high energy and no hydrogen bonds or helical units. This energy quickly decreases as a stabilizing bend in the structure is formed and hydrogen bonds first appear (structure G10). As the ϕ and ψ angles approach those of a helical residue, a hydrogen bond is lost, destabilizing the helical turn. Hydrogen bonds reform, however, allowing a stabilized semi-helical intermediate to persist for over two generations (structure G18). Soon after, the helical turn dissolves and a typical, solvated member of the unfolded ensemble emerges (structure G24). Referring to Figures 5 and 6, it is at approximately generation 20 that this folding trajectory becomes similar in nature to the folding pathway followed in other Series: an unfolded globule persists for some time, followed by a sudden plummet of exposed core surface area and the onset of hydrophobic contacts, leading to a folded hairpin later in

Figure 4. A detailed analysis of a trajectory resulting in the fully formed hairpin (Series 2). (a) A representation of the folding trajectory; the backbone of the peptide is represented as a cyan trace; the core hydrophobic residues (Trp43, Tyr45, Phe52 and Val54) are shown according to our previously defined color scheme. (b) RMSD from the 1GB1 structure of the hairpin (residues 43-54), radius of gyration and the number of backbone-backbone hydrogen bonds. (c) Distance between key hydrogen bonding partners (green, Trp43-Val54; red, Tyr45-Phe52), and the minimum distance between Trp43 and Phe52 (black). Note that the minimum distance between Trp43 and Phe52 reaches its final value before the key hydrogen bonds are established. (d) Solvation energy ($E_{\text{solvation}}$), charge-charge energy (E_{charge}), and total potential energy *versus* time. The initial hydrophobic collapse of the unfolded peptide correlates with a sharp decrease in E_{total} , while the attainment of the final structure correlates with E_{total} reaching its final value. A significant deviation of E_{charge} and $E_{\text{solvation}}$ from their final value around generation 160 is correlated with the temporary breaking of the key Tyr45-Phe52 hydrogen bond. (e) A concise summary of the key events along the folding trajectory (color code: yellow, high; violet, low). HB- ij denotes the distance between the hydrogen bonding partners i and j ; min- pq denotes the minimum distance between residues p and q . Note that the establishment of the Trp43-Phe52 interaction (most likely due to hydrophobicity) is the earliest event of significance along the trajectory.

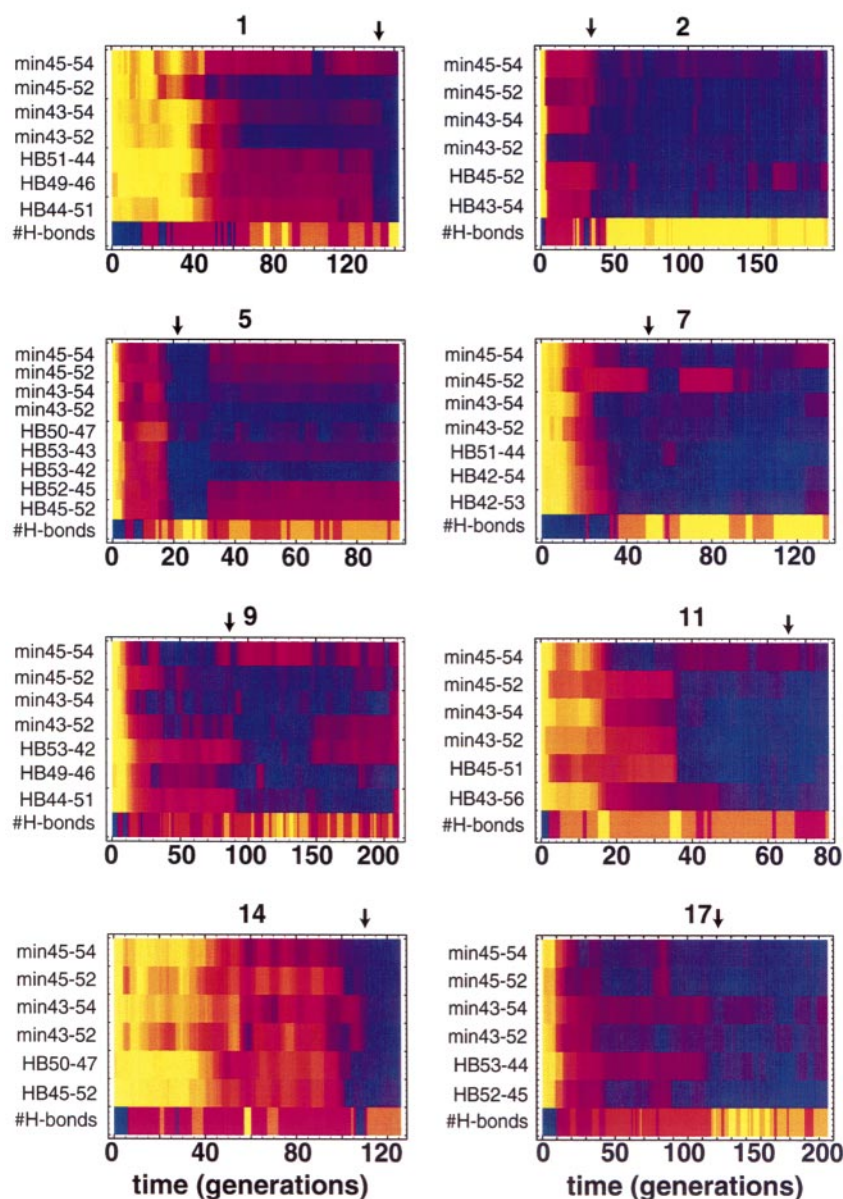


Figure 5. Graphical representation of successful folding trajectories (color code: yellow, high; violet, low). Each frame shows the total number of backbone-backbone hydrogen bonds (# H-bonds), the distance between key hydrogen bonding partners (HB- ij , where i and j stand for the sequence numbers of the residues), and the minimum distance between the core hydrophobic residues (min- pq , where p and q stand for the sequence numbers of the residues). The numbers above the frames indicate the internal coding numbers of the Series; time is expressed in the number of generations. All or most of the runs are characterized by an early appearance of hydrophobic interactions followed by the formation of hydrogen bonds connecting the two strands of the hairpin, and a fluctuating total number of hydrogen bonds. In each run, the first appearance of four or more residues with beta sheet dihedral angles is marked with an arrow. Several runs (5, 9 and 17) exhibited partial unfolding after the initial folding event. A semi-helical intermediate exists early in the Series 11 trajectory prior to generation 20 (refer to Figure 7 for an analysis of this event).

the trajectory. This semi-helical intermediate is encountered in only one of our eight successful trajectories. Figure 7 suggests that hydrogen bonding is responsible for stabilizing the semi-helical conformers witnessed herein. We have also examined the solvation energy and radius of gyration of the core throughout this event and neither shows any response to formation of the helical turn, supporting our assessment that hydrogen bonding plays the dominant role in stabilizing these intermediates.

In Figure 8 we plot the dependence of the statistical energy on the total number of hydrogen bonds and the minimum distance between the hydrophobic core residues, d_{\min} . The composite data set is again displayed in Figure 8(a) showing the trend of high hydrogen bonding following a decrease in core separation. However, it is clear that when our analysis includes all the unfolded

and partially folded structures this trend is rough at best. In Figure 8(b) the same parameters are examined for only the fully folded trajectories. Here there is one major “state” characterized by a narrow range of core separation values and a broad range of hydrogen bonds. In addition, there is a continuum of unfolded structures with varying values of d_{\min} , but a consistently low number of hydrogen bonds. In Figure 8(c) we plot the same profile based on the 15 ns equilibrium simulation of the hairpin from the 1GB1 structure of protein G.

Discussion

Understanding the folding of the β -hairpin is important, since parallel and antiparallel β -sheets are some of the most ubiquitous secondary

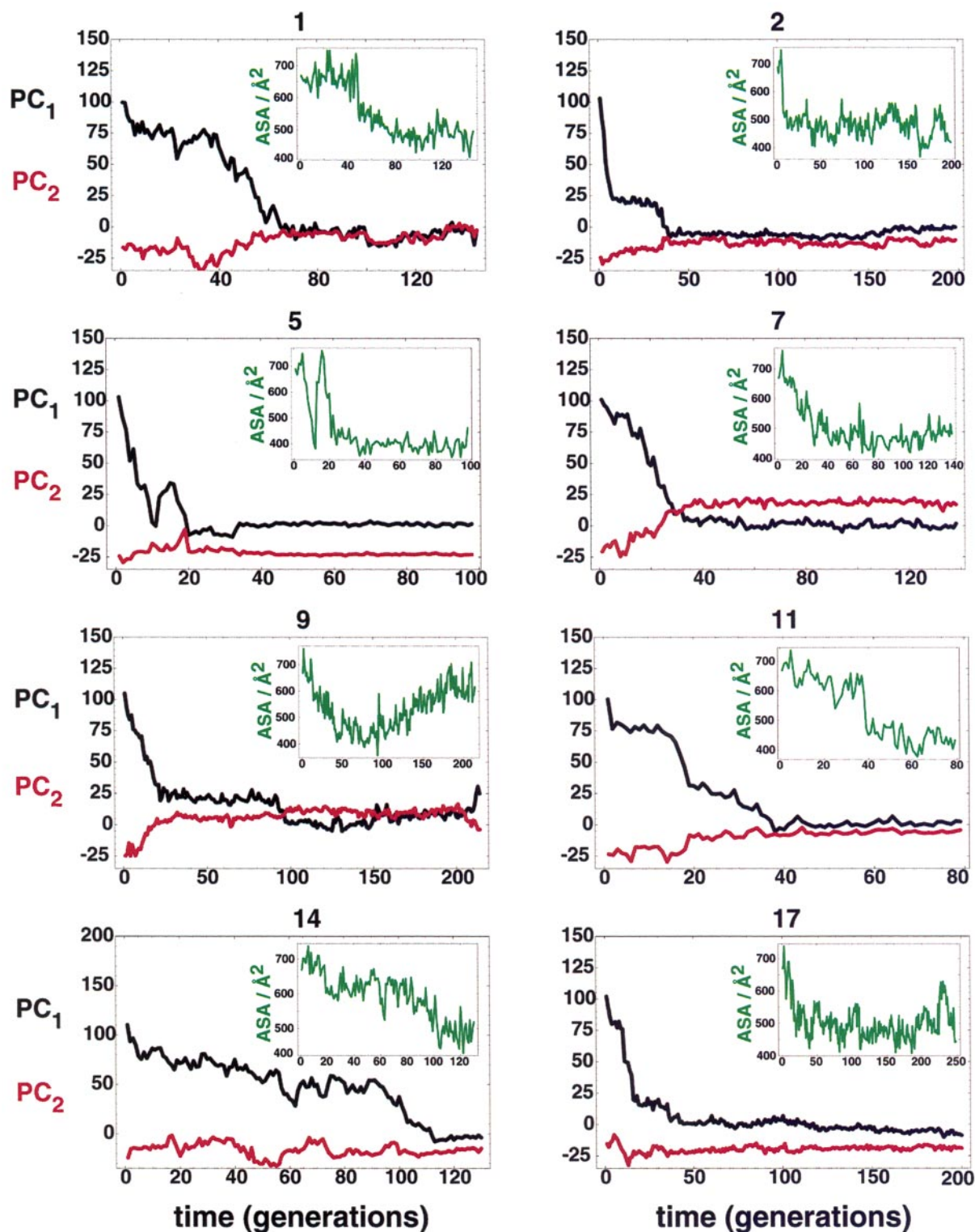


Figure 6. Characterizing hydrophobic collapse for successful folding trajectories. The time-dependent principal components PC₁ and PC₂ are plotted in black and red, respectively, for each of the eight independent successful runs. PC₁ should be regarded as the dominant folding mode throughout the folding event, and specifies the hydrophobic intermediate and folded conformations when approaching zero. PC₂, in comparison, characterizes the folding process quite poorly, revealing only the major transitions in configurational space. Insets: traces of the core solvent-accessible surface area (ASA) for each trajectory showing the rapid hydrophobic collapse that drives these folding events. Note the late decrease in ASA for Series 11 which is hindered early on by formation of a semi-helical intermediate conformation (see Figure 7).

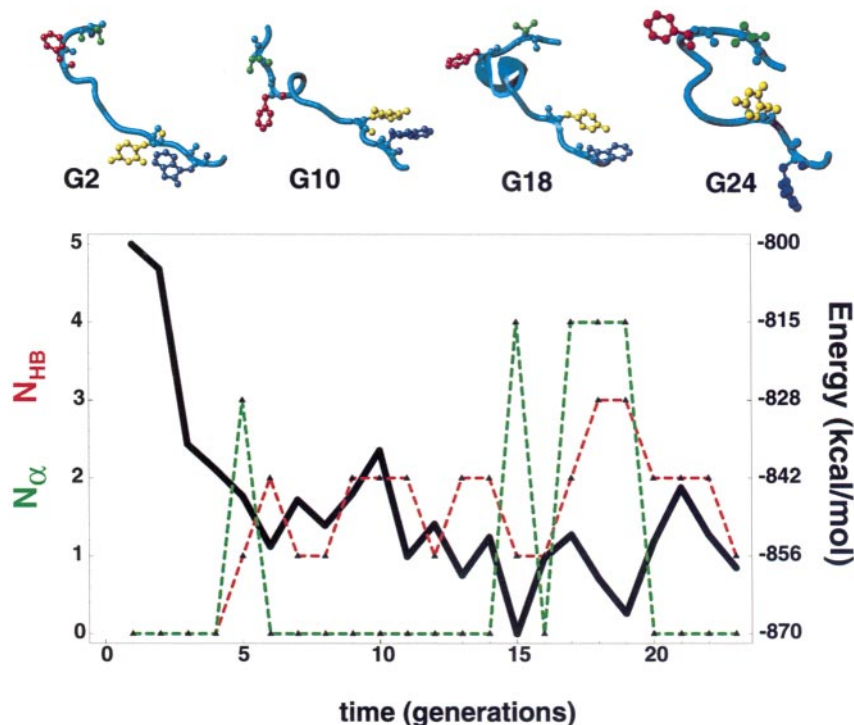


Figure 7. Analysis of the semi-helical intermediate from Series 11. Traces of the total potential energy, number of hydrogen bonds, and number of α -helical residues throughout the event in which an unfolded structure became semi-helical and then relaxed into a bent, solvated structure. Top: representations of the structures at specified generations. Note that semi-helical structures appear to be stabilized when hydrogen bonds are present ($G17 < t < G19$) but seem to rapidly dissociate when hydrogen bonds are absent ($G15$). This single transition through the semi-helical region of configurational space demonstrates well the short-lived nature of these meta-stable, semi-helical off-pathway intermediates.

structural elements in proteins. In this study we have reported atomistic, implicit solvent folding simulations of the hairpin at a biologically relevant temperature. Starting from the fully extended structure we have obtained a very large ensemble of conformations composed mostly of partially folded structures, as well as eight complete, fully independent folding trajectories. These data sets allow us to determine the key trends characterizing the folding process and determine several average properties that have been or could, in principle, be measured experimentally.

Several studies have characterized the folding of the C-terminal hairpin from protein G, and we wish to first compare these results to our own. Pande & Rokhsar reported the results of high-temperature unfolding and refolding simulations of the hairpin in which a discrete unfolding pathway was recognized to include a hydrophobically stabilized intermediate with only two to three hydrophobic contacts established in the core and little hydrogen bonding occurring.⁹ Qualitatively, the picture of the folding process from high-temperature unfolding presented in that study agrees well with the results of this study. The only significant difference between the two analyses concerns the role of water molecules in defining the partially solvated globular state, but given the implicit nature of the solvent in our simulations an equivalent state could not be identified. Garcia & Sanbonmatsu later verified the existence of the hydrophobic intermediate through a temperature-exchange Monte Carlo (MC)/molecular dynamics hybrid model of unfolding in which the thermodynamics of the unfolding events are well

described.¹⁷ This “H” intermediate was then observed in mechanically driven unfolding simulations performed by Bryant *et al.*,¹⁵ who describe it as including a nearly assembled core and very little backbone hydrogen bonding. However, Dinner *et al.*, who performed MC unfolding simulations on this peptide, report the 300 K energy landscape to be composed of a single descent from the unfolded state to the hairpin structure, and present a statistical energy plot that very closely resembles Figure 3(a).¹⁰

What then is the true energetic profile that describes this process? A look at the complete folding trajectories (Figure 5) offers a direct observation of the hydrophobically collapsed H intermediate previously reported. Furthermore, the statistical energy well characterized by low minimum core separation and zero to two residues with β -sheet backbone angles in Figure 3 clearly corresponds to the H state. In this sense the folding process can be described using a three-state scheme with a semi-helical off-pathway intermediate present: $I_x \rightleftharpoons U \rightleftharpoons H \rightleftharpoons F$. We believe that a large ensemble of complete folding trajectories would result in free energy plots which would clearly include all of these states as well as distinct barriers between them.

The absence of explicit solvent molecules in our simulations precluded us from analyzing in detail the role of solvation in folding of the hairpin. Nonetheless, we believe that the implicit generalized Born (GB) model of the solvent that we have used is sufficiently realistic and that the presence of explicit water molecules would not significantly affect our conclusions. Brooks and his collaborators

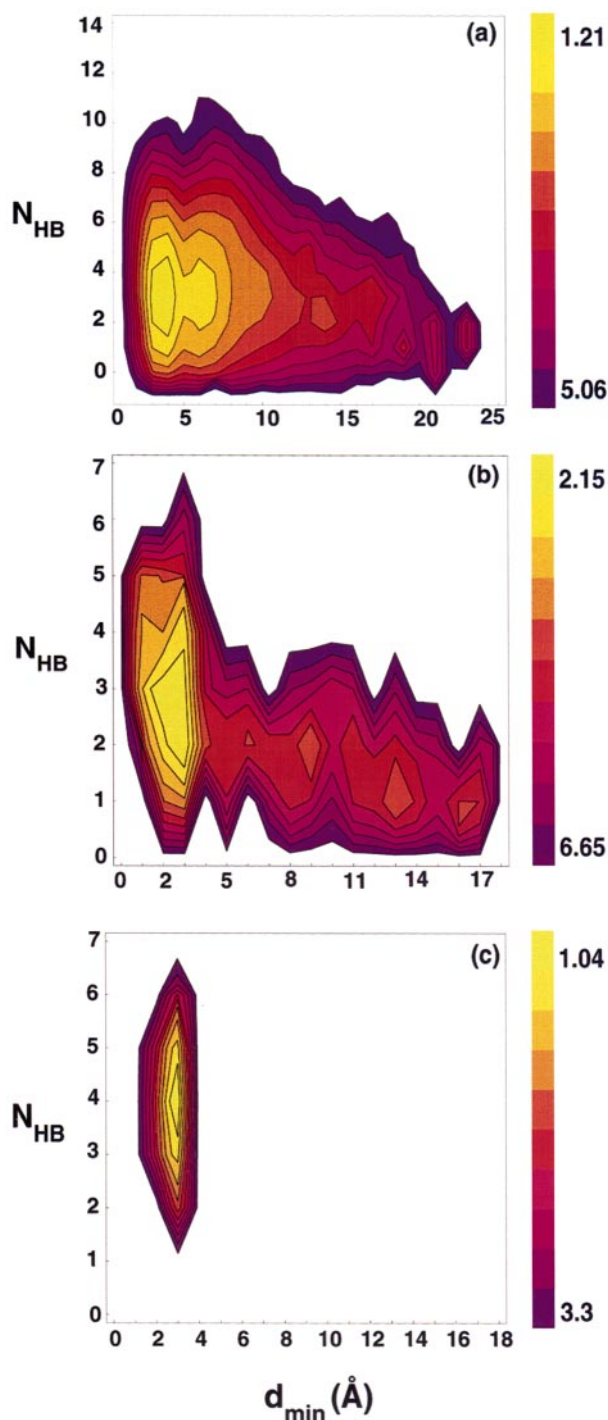


Figure 8. Relative statistical energy analysis for three data sets. The statistical energy as a function of the minimum distance between the hydrophobic core residues (see the text) and the total number of backbone to backbone hydrogen bonds for (a) all data, (b) the eight trajectories which folded into the final hairpin structure, and (c) the 15 ns equilibrium run of the 1GB1 structure of the hairpin. The contours are drawn at linearly increasing intervals between the lowest and the highest values shown with each scale bar.

have used molecular dynamics to study folding of the entire protein G, and have suggested an important role for water in “lubricating” conformational transitions of the collapsed intermediate.^{26–28} In their picture, water is squeezed out of the protein interior in the late stages of folding. We believe that the solvent plays a less important role in our case, primarily because of the size of the fragment that we have simulated. The hydrophobic core of the hairpin consists of just four residues, and even if the solvent were represented explicitly it probably would not significantly alter the solvation and desolvation of the core of the peptide. Bursulaya & Brooks²⁹ have used both the explicit and the implicit solvent model to look at folding of Betanova,³⁰ a three-stranded antiparallel β -sheet, and have shown that water does not play a significant role in its folding process. They concluded that the generalized Born implicit solvent model captures most aspects of solvation important for folding of Betanova, and have surmised that the same may hold for other small polypeptides. Finally, we note that numerous studies have successfully applied the generalised Born/surface area (GB/SA) model not just to proteins, but also to various biological systems including RNA and ligand-receptor binding pairs.^{31–33} While investigations of the overall accuracy of the GB/SA model in comparison to both experimental data and the solvation energies calculated using the much more rigorous Poisson-Boltzmann (PB) equation have found quantitative differences between GB and PB calculations for large macromolecules, these studies generally find that relative energies are well correlated and that the GB model approximates the true solvation energies quite well.^{34,35}

The occurrence of short-lived, semi-helical intermediates in our simulations seems at first hard to reconcile with experimental data on hairpin formation: the NMR and the CD measurements on the hairpin alone in solution by the Serrano group revealed no signature traces of α -helix.^{5,6} However, there are three important factors regarding our simulations which must be considered. First, though we see as many as nine helical residues within a single structure, these largely helical conformations ($N_\alpha > 5$) account for less than 0.4% of the entire ensemble of unfolded structures and have not been seen in trajectories leading to folded hairpin conformations. More importantly, while these highly helical structures are relatively persistent (witnessed to varying degrees for as long as ~ 10 ns with high variability in the total number of α -like residues), the more common semi-helical intermediates described herein last less than a nanosecond. These time-scales are not only very short relative to the overall folding time of the peptide, but also far too short to be observable experimentally. It should not be a surprise, in fact, that an unfolded peptide would randomly sample conformations with small helical portions stabilized by hydrogen bonds. Furthermore, our results agree

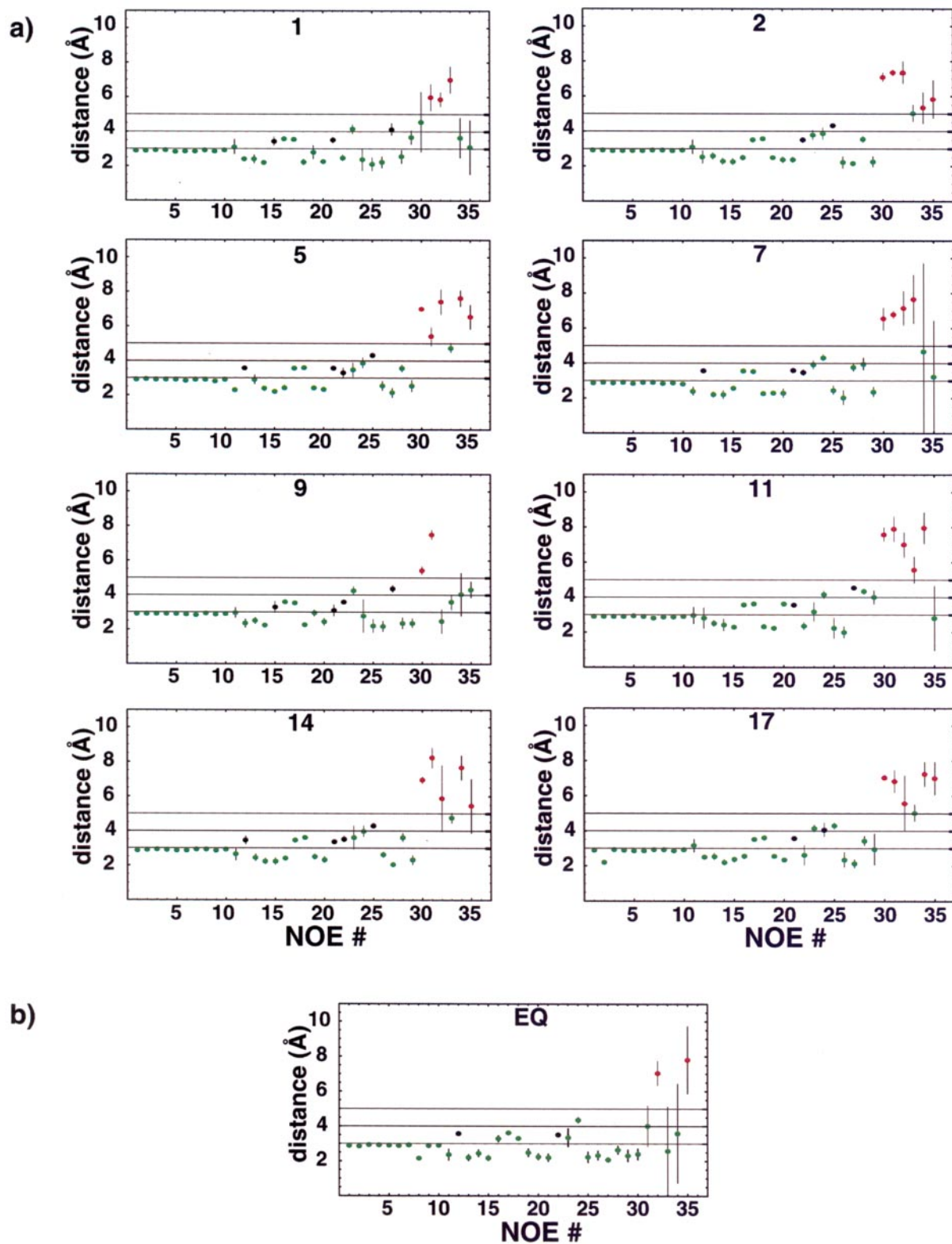


Figure 9 (legend opposite)

well with the description by Garcia *et al.*, which states that the observed helical structures are less favored energetically in comparison to the hairpin conformation.¹⁷ The semi-helical intermediate described in Figure 7, for example, is approxi-

mately 30 kcal/mol higher in energy than the final hairpin conformation from that run. At this point we should comment on the applicability of using the total potential energy as a measure of folding. While it is true that the helical intermediates we

saw were typically higher in total potential energy than the folded hairpins, we observed many unfolded structures in our composite data set which were comparable in energy to the hairpins. Also, while in our complete folding trajectories the energy was always at its minimum for the hairpin structures, there were also other structures of the same minimum energy which were not hairpin-like. In other words, in our study we could not use the total potential energy as a sole indicator of folding. This clearly poses a challenge if we are to extend the ensemble dynamics methodology to structure prediction. Nevertheless, we believe that regarding this problem the main area where improvements could and should be made is potential set design. The ensemble dynamics method, with its unparalleled ability to reach long simulation time-scales, could play a significant role in this process as well.

A striking aspect of our folded ensemble is the fact that, despite the well packed core and β -sheet backbone dihedral angles, different folded structures exhibit varying interstrand hydrogen bonding patterns which in turn differ from the hydrogen bonding seen in the NMR structure of protein G (1GB1).⁴ Since there is no experimental structure of the hairpin alone in solution, an important question that we have to address is this: how do we know that what we call "folded ensemble" really does represent the native, solvated β -hairpin? The major experimental results concerning the structure of the hairpin alone in solution come from NMR measurements by the Serrano group,^{5,6} and to a lesser extent from fluorescence

measurements by the Eaton group.⁷ The NMR measurements provided a set of NOE constraints that the folded structure should satisfy, and in Figure 9 we compare our final structures against these constraints. In short, we have calculated the distances between pairs of atoms which exhibit NOEs in the experiment and asked whether or not these distances fall within the distance ranges typical of the NOEs of the experimentally observed strength (2-3 Å for strong, 2-4 Å for medium and 2-5 Å for weak and very weak NOEs). Figure 9(a) shows the results for all of the unambiguous NOEs reported in the NMR study by the Serrano group,⁵ and the agreement with the experimental data is good. For all pairs except two (C^α - C^α 45-52 and C^α - C^α 43-52), we expect that the ensemble of folded structures obtained by simulation could, in principle, give rise to the experimentally observed NOEs. The fact that our simulations agree poorly with the weak NOEs between the C^α protons of the hydrophobic core residues merits attention. While all of our folded structures exhibit hydrophobic cores with closely packed side-chains, they all share an intriguing feature: their strands are slightly out of plane with each other, resulting in the higher than expected separation between the C^α protons of the core residues. This may be due to several factors including methodological ones such as slight imperfections in the force field, or physical ones such as flexibility in the turn region. In Figure 9(b) we show an analogous analysis for the 15 ns equilibrium trajectory started from the 1GB1 structure. Interestingly enough, after only a few nanoseconds of equilibration there are already

Figure 9. A comparison of the simulation results with the NMR measurements of the hairpin in solution.⁵ Since the OPLS potential set does not explicitly include all the hydrogen atoms present in the molecule, for the purpose of this analysis we have used X-PLOR⁴⁵ and Charmm 22 potential set⁴⁶ to add all missing hydrogen atoms. For each of the Series resulting in folded hairpin (1, 2, 5, 7, 9, 11, 14 and 17, Figure 9(a)), and for the equilibrium run starting from the experimental 1GB1 structure (EQ, Figure 9(b)), we compare the calculated interatomic distances between the relevant pairs of protons with all the unambiguous NOEs seen experimentally.⁵ Thirty-five proton pairs, numbered 1 through 35, are shown on the x -axis of each plot with the average distances (see below) between them on the y -axis. The proton pairs with the assigned NOE strengths from Figure 2 in Blanco *et al.*⁵ are (vw, very weak; w, weak; m, medium; s, strong): 1. H^α -NH E42 (w); 2. H^α -NH T44 (w); 3. H^α -NH Y45 (w); 4. H^α -NH D46 (w); 5. H^α -NH D47 (m); 6. H^α -NH A48 (m); 7. H^α -NH T49 (m); 8. H^α -NH K50 (m); 9. H^α -NH T55 (w); 10. H^α -NH E56 (w); 11. H^α G41-NH E42 (m); 12. H^α E42-NH W43 (s); 13. H^α T44-NH Y45 (m); 14. H^α Y45-NH D46 (s); 15. H^α D46-NH D47 (s); 16. H^α D47-NH A48 (m); 17. H^α A48-NH T49 (m); 18. H^α T49-NH K50 (m); 19. H^α T51-NH F52 (m); 20. H^α T53-NH V54 (s); 21. H^α V54-NH T55 (s); 22. H^α T55-NH E56 (s); 23. NH Y45-NH D46 (vw); 24. NH D46-NH D47 (vw); 25. NH D47-NH A48 (m); 26. NH A48-NH T49 (m); 27. NH T49-NH K50 (m); 28. NH K50-NH T51 (w); 29. NH T55-NH E56 (vw); 30. H^α Y45- H^α F52 (vw); 31. H^α W43- H^α V54 (vw); 32. H^α K50-3H Y45 (vw); 33. H^α Y45- 2H F52 (vw); 34. H^β Y45-5H F52 (vw), and 35. H_β W43-4H F52 (vw). The distances shown are averaged according to $R_{AVG} = \langle R_{HH}^{-6} \rangle^{-1/6}$ calculated over a 1 ns window after folding is complete. The error bars correspond to the variance around R_{AVG} . The distances corresponding to the equilibrium 1GB1 run are averages calculated in the same fashion over a 1 ns window between 3.5 and 4.5 ns after the beginning of the run. The points labeled in green match the experimentally observed NOE intensities (R_{AVG} is within 2-3 Å when strong NOEs are expected, 2-4 Å when medium NOEs are expected and 2-5 Å when weak NOEs are expected). The points labeled in black do not match the experimentally observed intensities but are still likely to be apparent in NOE spectra ($R_{AVG} < 5$ Å). The pairs of protons corresponding to the points labeled in red probably would not result in observable NOE signals ($R_{AVG} > 5$). For clarity, we show horizontal lines at 3 Å (the upper bound for strong NOEs), 4 Å (the upper bound for medium NOEs) and 5 Å (the upper bound for weak NOEs). Points 1-22 are largely independent of conformation, but we include them for completeness. Albeit crude, this analysis shows that our folded ensemble could in principle give rise to the experimentally observed NOEs. It is likely that a larger folded ensemble would be in an even better agreement with the experiment.

several proton pairs which, unlike in the starting structure, violate the distance constraints derived from the experiment. This suggests that methodological factors such as the force field are most likely responsible for any slight inconsistencies between the experimental results and our simulations. Comparing our results with the results of the fluorescence measurements has been much more difficult, since the distance constraints that could be extracted from fluorescence measurements are less detailed and less strict than their NMR counterparts. Nevertheless, we have calculated and/or estimated the distances between the fluorophores and quenchers used in the experiment⁷ (data not shown), and all of our folded structures fell below the characteristic distance $r_0 = 2.2$ nm used in the analysis of the experiment. All in all, the above analysis gives validity to our claim that the folded ensemble obtained in our simulations does represent the behavior and the structure of the folded β -hairpin in solution. The heterogeneity of the hydrogen bonding patterns that we observe in the folded ensemble can, after all, be expected from the solvated, stand-alone hairpin lacking all of the stabilizing tertiary interactions provided by the rest of the protein G molecule.

Based on our results we can estimate the folding rate of the hairpin in the following way: we have simulated 27 independent Series, each including 100 Trial simulations which on average complete approximately 14 ns of simulated time, bringing the total to approximately 38 μ s of real time simulation. Out of 2700 Trials, we have detected eight complete folding events, which if we assume single-exponential folding behavior, as witnessed experimentally, results in an estimated folding time of $4.7(\pm 1.7)$ μ s (see Methods for folding time calculation and error estimation). This estimate is in excellent agreement with the experimentally measured time of 6 μ s.⁷ Our value of $4.7(\pm 1.7)$ μ s is an upper bound on the folding rate, and would be exact if there were no coupling between the Trials in each run. However, since one half of the trajectories (four out of eight) within our folded ensemble never exhibited any 300 kcal²/mol² spikes in energy variance (meaning their Trials explored the phase space independently from the beginning to the end: there was no coupling between them), and the other half of the trajectories each went through just one early transition, this estimate is fairly accurate. If we include all the folding events that we have observed disregarding the fact that some Trial simulations were coupled to each other (there were 40 such events in all) we get the value of $0.9(\pm 0.15)$ μ s as our prediction of the lower bound on the folding rate. In another simulation study covering long time-scales, Ferrera & Caflisch have used an implicit solvent model based on the solvent-accessible surface area to analyze folding of a three-stranded β -sheet, and their estimated folding times were approximately two orders of magnitude faster than the experimentally observed values.³⁶ They accounted for this discre-

pancy by noting that their implicit solvent model may artificially smoothen the potential energy landscape traversed by the folding molecule and that they did not account for the friction exerted by the solvent molecules. We believe that the accuracy of the rate estimate seen in our study is due to the inclusion of the friction term in our description of the system as well as to the additional generalized Born term employed in the solvent model.

Our results offer the following picture of the folding process. Folding from a fully extended conformation begins with a rapid collapse to a more compact structure. During this time, different temporary hydrogen bonds form, condensing the peptide and decreasing the costly loop entropy which hampers the formation of the hydrophobic core. These temporary hydrogen bonds form and break and, in general, their pattern which varies from run to run has no resemblance to the final hydrogen bonding pattern of the hairpin. Next, an interaction between the hydrophobic core residues is established, and this is clearly the central event in the folding process and most probably its rate-limiting step. Note that at this point the core is still not fully formed: the initial hydrophobic interaction most often involves just two or at most three hydrophobic residues on the opposite sides of the future hairpin. Full formation of the core appears typically simultaneously with the establishment of final hydrogen bonds. The folded hairpin structure is characterized by a diverse pattern of fluctuating backbone hydrogen bonds, with two or more being fairly consistent, and a stable and well packed core. This structure of the folded state is consistent with the explicit solvent molecular dynamics studies of the folded state of the molecule by Ma & Nussinov.¹⁶ They showed that hydrogen bonds are not at all critical for keeping the hairpin together, ascribing the stability of the hairpin to the hydrophobic core. Also, Berendsen and co-workers¹¹ have shown using similar methods that the main equilibrium motions of the peptide involve large fluctuations in the turn and end regions with the core remaining stable during the simulations.

The turn region of the peptide (residues 46-51) also merits some attention. The folded ensemble from our simulations exhibits significant variability in this region when it comes to hydrogen bonding pattern and side-chain packing. Furthermore, each particular trajectory from the folded ensemble exhibits significant fluctuations in this region even after folding is complete. This is not inconsistent with previous experimental and theoretical results. The 3D structures of protein G obtained by different experimental means (NMR^{4,22} and X-ray^{20,21}) differ significantly from each other in the turn region, suggesting it may be quite flexible. Also, an equilibrium molecular dynamics study of the solvated hairpin¹¹ showed significant flexibility in this region, confirming our results. Analyzing the effects of mutations on folding kinetics of protein G, McCallister *et al.*³⁷ have shown that truncation

of Asp46 in the turn region to Ala, which removes a hydrogen bond between this residue and Ala48, slows the folding rate 20-fold with no effect on the unfolding rate. We have analyzed our data with a focus on this particular hydrogen bond, but no significant patterns were observed. In five out of eight members of the folded ensemble this hydrogen bond appeared temporarily before folding was complete, but we cannot say to what extent or if at all it had any influence on the folding mechanism.

Our picture of the folding process is, in essence, a blend of the hydrogen bond-centric and the hydrophobic core-centric views of the folding of the hairpin: non-specific hydrogen bonds are important in the initial stages of folding, and are capable of stabilizing both semi-helical intermediates and hairpin precursors alike, but the key event which stabilizes the bent precursor of the hairpin and guides the downstream folding process is the formation of a hydrophobic interaction between the core residues. Final hydrogen bonds appear later, around the same time the full formation of the hydrophobic core occurs, and these continue to fluctuate even after folding is complete. This last fact is well demonstrated by the statistical energy plot in Figure 8(b), which shows that the folded state is characterized by large fluctuations in the total number of hydrogen bonds. Such behavior is fully consistent with the equilibrium behavior of the hairpin from the 1GB1 structure (Figure 8(c);^{11,15,16}). Pande & Rokhsar,⁹ Dinner *et al.*¹⁰ and Bryant *et al.*¹⁵ have used different simulation techniques to look at unfolding of the hairpin, and their results are in fair agreement with ours: they also emphasize the major role of hydrophobic interactions in the early formation of the hairpin. On the other hand, our results disagree with a proposed mechanism^{7,8} in which folding initiates at the turn and propagates down the hairpin in a zipper-like mechanism. One potential source of discrepancy is the difference between models: the zipping model proposed by Eaton *et al.*^{7,8} by construction excludes the possibility of folding being initiated by the formation of the hydrophobic core. Skolnick and co-workers¹² have used a lattice model of protein structure and dynamics to look at hairpin formation and also predominantly saw the turn form first, in agreement with the results from the Eaton group. Less frequently, they also witnessed the core formation occur first. Note, however, that, in our model, a turn-like structure forms first as well: after all, a hydrophobic interaction between residues on opposite sides of the hairpin has to result in the formation of a turn by necessity. The overall coarse grained nature of the Skolnick model may have resulted in their identifying the partial hydrophobic core formation as formation of the turn. Finally, our main conclusions are supported by recent ²H/¹H amide kinetic isotope effect experiments on ubiquitin by Sosnick and co-workers.³⁸ They observed that in order for ubiquitin to

achieve the correct β -sheet topology, extensive native hydrogen bonding may not be required.

A long-standing question in the field of protein folding concerns the applicability of different potential sets for modeling proteins. The results of our study generally speak in favor of using the OPLS potential set³⁹ for studying folding of protein fragments and small proteins. However, the possibility that this set is not perfectly suitable for the direct folding of the β -hairpin remains. "Helix-friendliness" of potential sets might provide an explanation for the helical structures now seen in both molecular dynamics (MD) and MD/MC hybrid simulations that have not been established experimentally. This may not be the case, of course, if these less stable helical structures truly are short-lived (nanosecond regime) as predicted in our model. We look forward to experimental verification of the existence of this ensemble in an aqueous environment at biologically relevant temperatures.

Conclusion

To address the important questions posed regarding β -hairpin folding, we have made the following conclusions based on both the composite and fully folded data sets described herein: (a) the folding pathway includes a hydrophobically stabilized "H" intermediate, characterized by two to three core contacts and little backbone hydrogen bonding; (b) an off-pathway semi-helical intermediate state (trap) is possible early on in the folding process of this peptide; (c) partial hydrophobic core formation takes precedence over interstrand hydrogen bonding as the important interaction to initiate folding; (d) although temporary, non-specific hydrogen bonding is involved in early collapse of the peptide, the final backbone hydrogen bonding pattern tends to form in tandem with complete core formation; and (e) among the folded hairpin structures, no specific backbone hydrogen bonding pattern is prevalent, and a sampling of many possible schemes is expected at equilibrium.

This study suggests the great potential of using distributed computing paradigms and computer mega-clusters to study protein folding. We have simulated in fine detail more folding time than has ever been reported for a peptide of this size: this has allowed us to directly observe and analyze folding events which take place on the microsecond time-scale. The results we obtained agree well with current experimental and theoretical knowledge of hairpin folding, speaking in favor of the presented methodology. As the ensemble dynamics method develops, we hope it will lead not only to answering many questions of immediate biological importance, but also to improving force fields, implicit and explicit solvent models and biomolecular computation in general.

Methods

We have simulated in atomistic detail the folding of the β -hairpin from the C terminal part of the immunoglobulin (Ig)-binding streptococcal protein G (PDB code 1GB1, residues 41-56). The capped sequence of the peptide is Ace-GEWTYDDATKTFVTE-NH₂. The folding of the peptide was modeled using the OPLS parameter set,³⁹ the GB/SA implicit solvent model³⁵ and the molecular dynamics software from the TINKER molecular modeling package[†]. Langevin dynamics was used to simulate the viscous drag of water ($\gamma=91$ ps⁻¹), the bond lengths were constrained using the RATTLE algorithm,⁴⁰ and a 2 fs integration time step was used. For electrostatic calculations, 16 Å cutoffs with 12 Å tapers were employed. The temperature was held constant at 300 K.

A distributed-computing/ensemble-dynamics approach²³⁻²⁵ was used for the simulations and involved utilizing a supercluster of processors around the world. The crux of this dynamics method is as follows:[‡] Many complex phenomena including protein folding are modeled as a system crossing multiple free energy barriers. Since the bulk of simulation time is spent exploring the free energy wells and waiting for thermal fluctuations to bring the system across the barriers, one can speed the calculations by starting multiple independent simulations in the first free energy well and waiting for one simulation to cross the first barrier. At this point the simulations are coupled by transferring them all to the configuration space location of the one which has crossed and all simulations are independently restarted from there. This process is repeated as many times as needed to traverse the remaining statistical energy barriers.

Such a scheme can achieve an M -times simulation speedup using M processors.²⁴ Consider the single barrier case. For a single processor, we expect that a simulation would have crossed the barrier in time t with a probability:

$$P_1(t) = k \exp(-kt)$$

For the M -simulation case, the probability that the first simulation has crossed in time t is:

$$P_M(t) = [k \exp(-kt)]M \left[1 - \int_0^t k \exp(-kt) \right]^{M-1}$$

(i.e. the product of the probability that one simulation has crossed, times a degeneracy factor of M , since we do not care which simulation crossed first, times the probability that the remaining $M - 1$ simulations have not folded). Evaluation of the integral above yields:

$$P_M(t) = Mk \exp[-Mkt]$$

which is exactly the same distribution as the single processor case, with an M times increase in the effective rate. Since this method speeds transitions for each barrier crossing by a factor of M , it will also speed the multiple barrier case by M times.

We have applied this paradigm for looking at the folding of the β -hairpin. Folding is initiated from a fully extended conformation on $M = 100$ different Trial processors (which defines one Series, see Figure 10), each processor initiated with a different random number seed.

If, and when, one of the Trial simulations crosses a free energy barrier (i.e. exhibits a state transition, as defined by a spike of at least 300 kcal²/mol² in the energy variance time-series) all other Trials in that particular Series are transferred to the configuration-space location of the one that has just made the transition, and all are independently restarted from this point with new random number seeds. Note that if no transitions are detected, the ensemble dynamics scheme reduces to running a large number of fully independent simulations in parallel. In order to reconstruct folding trajectories for those runs which exhibited state transitions (and whose Trials were therefore mutually coupled), we concatenate the fragments of those Trial trajectories which lead to the observed transitions. We present 27 independent Series, each one including 100 Trials (Figure 10), for a total simulation time of approximately 38 μ s. The structures used in the analysis were recorded every 100 ps.

We define a structure as being “folded” if it meets the following criteria: (a) the backbone dihedral angles (ϕ, ψ) have adopted proper β -sheet values; (b) the hydrophobic core is well packed; and (c) multiple hydrogen bonds across the hairpin channel are present. We do not refer to “native” hydrogen bonds, unless explicitly specified, since a diversity of final hydrogen bonding patterns has been witnessed.

To analyze the trajectories obtained, ten degrees of freedom were chosen to characterize the folding landscape and kinetics of the hairpin. The number of hydrogen bonds, N_{HB} , was defined as the number of carbonyl-amide pairs separated by 2.5 Å or less with a minimal chain distance of three residues. The radius of gyration, R_g , was calculated based on all heavy atoms and accounts for atomic masses. As an additional geometric parameter, we define the minimum core separation, d_{min} , as the smallest spacing between core residue side-chain atoms (C α excluded) located on opposite strands of the hairpin. This measurement ignores the hydrophobic packing of two core residues on the same side of the hairpin, and can thus aid in distinguishing configurations in which the core is properly formed from those in which the backbone is bent in a hairpin-like manner yet exhibit improper side-chain orientations and/or a strong degree of asymmetry about the β -turn. To address the overall folding scheme of each trajectory individually, the time-dependence of the first two principal components (PC₁ and PC₂) was examined in each case. Principal component analysis has been well described^{17,41,42} so we include a very brief review. Eigenvectors were found for the symmetric matrix given by $\sigma_{\alpha\beta} = \langle \Delta r_\alpha \Delta r_\beta \rangle$, where Δr_i is the displacement from the i th average atomic coordinate. Each conformation was then projected onto the eigenvectors that generated the two greatest eigenvalues in the equation $\sigma_{\alpha\beta} \hat{e} = \lambda \hat{e}$, thus giving a least-squares style fitting of the majority of molecular motions observed to the eigenvectors obtained. These projections are the magnitudes of the principal components detailed herein. We present very good approximations to the exact principal components by including solely the 85 carbon atom positions, resulting in the reduced $3N = 255$ dimension configurational space. In our model, with folded states located at PC₁ ≈ 0 , this parameter represents the variation of a given conformation from the folded state. Though this single parameter does not include all motions within the hairpin, it does include the important motions along the folding trajectory (as described in previous sections), and should be intuitively regarded as the dominant folding mode of this process. Additionally, the total potential

[†] Available online at <http://dasher.wustl.edu/tinker/>

[‡] Available online at <http://folding.stanford.edu>

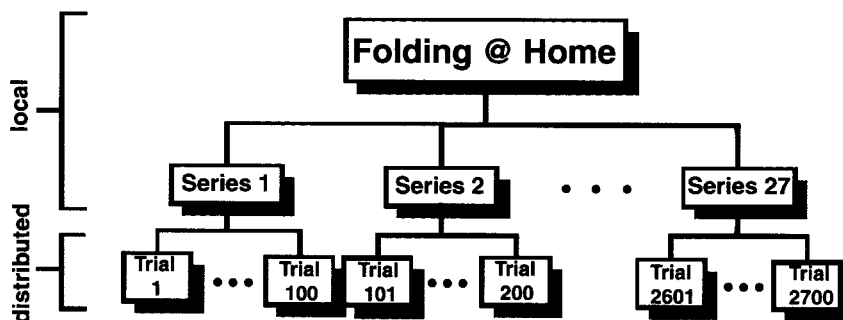


Figure 10. A schematic representation of the ensemble dynamics methodology. Locally, the Folding@Home servers initiate numerous Series of simulations. Each Series is independent of all others, and consists of 100 Trial simulations which are distributed to our thousands of users around the world. The data presented here were collected using a total of 27 Series offering 2700 Trials which resulted in a total of 38 μ s of simulated time eight fully independent, successful folding trajectories. The Trials are coupled within a single Series as follows: when a given Trial crosses a free energy barrier (herein defined by an energy variance of 300 kcal²/mol²) all other Trials within that Series are restarted from that configuration, all 100 having different random number seeds (and thus different random force components) upon each restart. If no transitions above the required energy variance are detected, the method simplifies to a mass parallelization of fully independent simulations.

energies and ASA were calculated using the “analyze” and “spacefill” routines in TINKER version 3.8. Energy calculations included the GB/SA solvent term which employs the analytical Still method for calculating polarization energies and included a surface area prefactor of $\sigma = 4.9$ kcal/mol². The surface areas were found using a 1.4 Å probe sphere radius and included hydrogen atoms.

We used two academic applications in our analysis. STRIDE^{43†} was used to determine the population of helical and sheet-like residues in each configuration, N_α and N_β respectively. This program considers both hydrogen bond energies and dihedral angles associated with a given set of atomic coordinates when assigning secondary structure to each residue in the sequence. To calculate RMS distances from the “native” hairpin (1GB1) ProFit[‡] was used. This code both aligns structures and minimizes the RMSD during calculations *via* translation and rotation of the structure alongside the templates. All of the RMSD values reported here are calculated for the α -carbon atoms of residues 43-54, since the two residues at each end of the hairpin are frayed in solution.

Based on our data we have estimated the folding rate and time constant in the following way. We assume (as seen experimentally) that the folding of the hairpin 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:

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

In the case of an ensemble of mutually 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 which 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:

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

If we assume that all of our 2700 Trial simulations ran independently (see Discussion for details), this means that in 14 ns, which is the time span covered by each Trial simulation, we have managed to fold 8/2700 or ~ 0.003 of the entire folding ensemble, resulting in the rate $K = 2.1 \times 10^5$ s⁻¹ or the time constant $\tau = 4.7$ μ s. Since (see Discussion for details) some of our Trial simulations were coupled with each other, this number is an upper bound on the folding rate.

The error inherent in the above procedure for calculating the rate and the time constant can be estimated in the following way. As shown above, for a given N_{total} and a given t , the rate K is simply proportional to the number of molecules that have folded by time t , $N_{\text{folded}}(t)$. Since each folding process behaves probabilistically (according to exponential distribution), then, given fixed t and N_{total} , the number of processes which will fold by time t , $N_{\text{folded}}(t)$, will be a random variable. In other words, different realizations of the “large experiment” containing N_{total} individual processes will, by their very nature, yield different values of $N_{\text{folded}}(t)$ for a fixed time t . From this it follows that our rate estimate will also be associated with a certain inherent uncertainty. From elementary probability theory, we know that the number of folding events by time t , $N_{\text{folded}}(t)$, given a constant rate, will be distributed according to the Poisson distribution.⁴⁴ This in turn means that the rate estimate, which is proportional to $N_{\text{folded}}(t)$, will also be distributed according to the Poisson distribution. The standard deviation of a Poisson distribution with rate λ is equal to $\lambda^{1/2}$, meaning that our rate estimate \pm standard deviation will simply be:

$$K = N_{\text{folded}}/(N_{\text{total}}t) \pm N_{\text{folded}}^{1/2}/(N_{\text{total}}t)$$

Standard propagation of error results in time constant

† Available online at <http://www.embl-heidelberg.de/stride/stride.html>

‡ Available online at <http://www.biochem.ucl.ac.uk/~martin/swreg.html>

\pm standard deviation of:

$$\tau = 1/K = (N_{\text{total}}t)/N_{\text{folded}} \pm (N_{\text{total}}t)/N_{\text{folded}}^{3/2}$$

Using $N_{\text{total}} = 2700$, $N_{\text{folded}} = 8$, and $t = 14$ ns results in $K = 2.1 \times 10^5 (\pm 0.74 \times 10^4) \text{ s}^{-1}$, and $\tau = 4.7 (\pm 1.7) \mu\text{s}$.

Acknowledgments

We gratefully thank the thousands of Folding@Home contributors, who made this work possible. A complete list of contributors can be found at <http://FoldingAtHome.Stanford.edu>. This work was supported, in part, by the ACS-PRF 36028-AC4 grant and CPIMA seed funds. B.Z. was supported by an HHMI predoctoral fellowship. We thank Robert Baldwin, Tanya Raschke and all the members of the Pande group for useful discussions. Finally, we extend our appreciation to the reviewers of this manuscript for their useful questions and comments.

References

- 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.
- Dill, K. A. & Chan, H. S. (1997). From Levinthal to pathways to funnels. *Nature Struct. Biol.* **4**, 10-19.
- Dobson, C. M., Sali, A. & Karplus, M. (1998). Protein folding: a perspective from theory and experiment. *Angew. Chem. Int. Edit.* **37**, 868-893.
- Gronenborn, A. M., Filpula, D. R., Essig, N. Z., Achari, A., Whitlow, M., Wingfield, P. T. *et al.* (1991). A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. *Science*, **253**, 657-661.
- Blanco, F. J., Rivas, G. & Serrano, L. (1994). A short linear peptide that folds into a native stable beta-hairpin in aqueous solution. *Nature Struct. Biol.* **1**, 584-590.
- Blanco, F. J. & Serrano, L. (1995). Folding of protein G B1 domain studied by the conformational characterization of fragments comprising its secondary structure elements. *Eur. J. Biochem.* **230**, 634-649.
- Munoz, V., Thompson, P. A., Hofrichter, J. & Eaton, W. A. (1997). Folding dynamics and mechanism of beta-hairpin formation. *Nature*, **390**, 196-199.
- Munoz, V., Henry, E. R., Hofrichter, J. & Eaton, W. A. (1998). A statistical mechanical model for beta-hairpin kinetics. *Proc. Natl Acad. Sci. USA*, **95**, 5872-5879.
- Pande, V. S. & Rokhsar, D. S. (1999). Molecular dynamics simulations of unfolding and refolding of a beta-hairpin fragment of protein G. *Proc. Natl Acad. Sci. USA*, **96**, 9062-9067.
- Dinner, A. R., Lazaridis, T. & Karplus, M. (1999). Understanding beta-hairpin formation. *Proc. Natl Acad. Sci. USA*, **96**, 9068-9073.
- Roccatano, D., Amadei, A., Di Nola, A. & Berendsen, H. J. (1999). A molecular dynamics study of the 41-56 beta-hairpin from B1 domain of protein G. *Protein Sci.* **8**, 2130-2143.
- Kolinski, A., Ilkowski, B. & Skolnick, J. (1999). Dynamics and thermodynamics of beta-hairpin assembly: insights from various simulation techniques. *Biophys. J.* **77**, 2942-2952.
- Honda, S., Kobayashi, N. & Munekata, E. (2000). Thermodynamics of a beta-hairpin structure: evidence for cooperative formation of folding nucleus. *J. Mol. Biol.* **295**, 269-278.
- Kobayashi, N., Honda, S., Yoshii, H. & Munekata, E. (2000). Role of side-chains in the cooperative beta-hairpin folding of the short C-terminal fragment derived from streptococcal protein G. *Biochemistry*, **39**, 6564-6571.
- Bryant, Z., Pande, V. S. & Rokhsar, D. S. (2000). Mechanical unfolding of beta-hairpin using molecular dynamics. *Biophys. J.* **78**, 584-589.
- Ma, B. & Nussinov, R. (2000). Molecular dynamics simulations of a beta-hairpin fragment of protein G: balance between side-chain and backbone forces. *J. Mol. Biol.* **296**, 1091-1104.
- Garcia, A. E. & Sanbonmatsu, K. Y. (2001). Exploring the energy landscape of a beta hairpin in explicit solvent. *Proteins: Struct. Funct. Genet.* **42**, 345-354.
- Eastman, P., Gronbach-Jensen, N. & Doniach, S. (2001). Simulation of protein folding by reaction path annealing. *J. Chem. Phys.* **114**, 3823-3841.
- Eaton, W. A., Munoz, V., Hagen, S. J., Jas, G. S., Lapidus, L. J., Henry, E. R. & Hofrichter, J. (2000). Fast kinetics and mechanisms in protein folding. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 327-359.
- Derrick, J. P. & Wigley, D. B. (1994). The third IgG-binding domain from streptococcal protein G. An analysis by X-ray crystallography of the structure alone and in a complex with Fab. *J. Mol. Biol.* **243**, 906-918.
- Gallagher, T., Alexander, P., Bryan, P. & Gilliland, G. L. (1994). Two crystal structures of the B1 immunoglobulin-binding domain of streptococcal protein G and comparison with nmr. *Biochemistry*, **33**, 4721-4729.
- Lian, L. Y., Derrick, J. P., Sutcliffe, M. J., Yang, J. C. & Roberts, G. C. (1992). Determination of the solution structures of domains II and III of protein G from streptococcus by 1 h nuclear magnetic resonance. *J. Mol. Biol.* **228**, 1219-1234.
- Voter, A. F. (1998). Parallel replica method for dynamics of infrequent events. *Phys. Rev. ser. B*, **57**, 13985-13988.
- Shirts, M. R. & Pande, V. S. (2001). Mathematical analysis of coupled parallel simulations. *Phys. Rev. Letters*, **86**, 4983-4987.
- Shirts, M. R. & Pande, V. S. (2001). Screensavers of the world, unite! *Science*, **290**, 1903-1904.
- Sheinerman, F. B. & Brooks, C. L. (1998). Calculations on folding of segment B1 of streptococcal protein G. *J. Mol. Biol.* **278**, 439-456.
- Sheinerman, F. B. & Brooks, C. L. (1998). Molecular picture of folding of a small α/β protein. *Proc. Natl Acad. Sci. USA*, **95**, 1562-1567.
- Sheinerman, F. B. & Brooks, C. L. (1997). A molecular dynamics simulation study of segment B1 of protein G. *Proteins: Struct. Funct. Genet.* **29**, 193-202.
- Bursulaya, B. D. & Brooks, C. L. (2000). Comparative study of the folding free energy landscape of a three-stranded beta-sheet protein with explicit and implicit solvent models. *J. Phys. Chem. B*, **104**, 12378-12383.
- Bursulaya, B. D. & Brooks, C. L. (1999). Folding free energy surface of a three-stranded beta-sheet protein. *J. Am. Chem. Soc.* **121**, 9947-9951.
- Williams, J. D. & Hall, K. B. (2000). Experimental and computational studies of the G[UUCG]C RNA tetraloop. *J. Mol. Biol.* **297**, 1045-1061.

32. Williams, J. D. & Hall, K. B. (1999). Unrestrained stochastic dynamics simulations of the UUCG tetra-loop using an implicit solvation model. *Biophys. J.* **76**, 3192-3205.
33. Zou, X., Sun, Y. & Kuntz, I. D. (1999). Inclusion of solvation in ligand binding free energy calculations using the generalized-born model. *J. Am. Chem. Soc.* **121**, 8033-8043.
34. David, L., Luo, R. & Gilson, M. K. (2000). Comparison of generalized Born and Poisson models: energetics and dynamics of HIV proteins. *J. Comput. Chem.* **21**, 259-309.
35. 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.
36. Ferrera, P. & Caflisch, A. (2000). Folding simulations of a three-stranded antiparallel β -sheet peptide. *Proc. Natl Acad. Sci. USA*, **97**, 10780-10785.
37. McCallister, E. L., Alm, E. & Baker, D. (2000). Critical role of beta-hairpin formation in protein G folding. *Nature Struct. Biol.* **7**, 669-673.
38. Krantz, B. A., Moran, L. B., Kentsis, A. & Sosnick, T. R. (2000). D/H amide kinetic isotope effects reveal when hydrogen bonds form during protein folding. *Nature Struct. Biol.* **7**, 62-71.
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. Andersen, H. C. (1983). Rattle: a "velocity" version of the Shake algorithm for molecular dynamics calculations. *J. Comput. Phys.* **52**, 24-34.
41. Garcia, A. E., Blumenfeld, R., Gerhard, H. & Krumhansl, J. A. (1997). Multi-basin dynamics of a protein in a crystal environment. *Physica D*, **107**, 225-239.
42. Garcia, A. E. (1992). Large-amplitude nonlinear motions in proteins. *Phys. Rev. Letters*, **68**, 2696-2699.
43. Frishman, D. & Argos, P. (1995). Knowledge-based protein secondary structure assignment. *Proteins: Struct. Funct. Genet.* **23**, 566-579.
44. Karlin, S. & Taylor, H. M. (1975). *A First Course in Stochastic Processes*, Academic Press, San Diego.
45. Brünger, A. T. (1992). *X-PLOR, Version 3.1: A System for X-ray Crystallography and NMR*, Yale University Press, New Haven.
46. Brooks, B. R., Brucoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). Charmm: a program for macromolecular energy minimization and dynamics calculations. *J. Comput. Chem.* **4**, 187-217.

Edited by F. Cohen

(Received 20 April 2001; received in revised form 15 August 2001; accepted 15 August 2001)