

First-Principles Calculation of the Folding Free Energy of a Three-Helix Bundle Protein

Erik M. Boczko* and Charles L. Brooks III†

The folding and unfolding of a three-helix bundle protein were explored with molecular-dynamics simulations, cluster analysis, and weighted-histogram techniques. The folding-unfolding process occurs by means of a "folding funnel," in which a uniform and broad distribution of conformational states is accessible outside of the native manifold. This distribution narrows near a transition region and becomes compact within the native manifold. Key thermodynamic steps in folding include initial interactions around the amino-terminal helix-turn-helix motif, interactions between helices I and II, and, finally, the docking of helix III onto the helix I-II subdomain. A metastable minimum in the calculated free-energy surface is observed at approximately 1.5 times the native volume. Folding-unfolding thermodynamics are dominated by the opposing influences of protein-solvent energy, which favors unfolding, and the overall entropy, which favors folding by means of the hydrophobic effect.

Recent years have seen a significant advance in our understanding of the protein-folding problem—the mapping of a protein sequence to a three-dimensional structure. A hallmark of this progress has been the macroscopic characterization of folding intermediates for several proteins (1). Theoretical models have been constructed to explain the experimental behavior of these states (2). However, molecular-level information concerning structure, dynamics, and thermodynamic properties of proteins as they fold has been lacking from both theoretical and experimental studies.

Theoretical models can provide insights into important aspects of protein folding. It is now possible to simulate the entire folding-unfolding process for simple proteins with lattice models (3). Lattice studies can address basic questions about folding mechanism (4), such as whether general mechanisms like the framework model (5), the diffusion collision model (6), hydrophobic collapse (7), or folding funnels (8) are prominent in folding. These methods also yield predictions of the native states of simple proteins from their sequences alone (9). However, the simple representation of protein amino acids as single beads (or sometimes two, one to represent the main chain and one for the side chain) and the inability to incorporate solvent explicitly preclude lattice-model simulations from probing molecular details of folding and of folding intermediates.

The simulation of the protein-folding process in its entirety with all-atom models of protein and solvent has, to date, been computationally too demanding. However, these models have been used to study pro-

tein unfolding under conditions of high temperature, when unfolding is rapid (10), and to characterize isolated nonnative equilibrium states of proteins near the native conformation (11). Additionally, all-atom models have been used to investigate the properties of peptides and secondary structural elements with an implicit "divide and conquer" strategy (12).

To move beyond current studies based on all-atom models with explicit solvent, we have developed a synthesis of large-scale molecular simulations and clustering methods to construct a free-energy surface for the folding-unfolding process of a small three-helix bundle protein. Our focus is, therefore, on the thermodynamics of the folding-unfolding process and the relations between structure, dynamics, and energy along a thermodynamic folding-unfolding coordinate. We do not address issues of folding kinetics.

We examined the unfolding and folding of a three-helix bundle because it is one of the simplest folding motifs. Specifically, we examined a 46-residue subsequence from fragment B of staphylococcal protein A (13). This protein is one of a few that are small and have a well-defined secondary and tertiary structure, and whose folding is not complicated by the presence of disulfide bonds, cofactors, or metals. The native structure of this bacterial cell wall protein (14) has been determined both by x-ray crystallography (15) and nuclear magnetic resonance (NMR) spectroscopy (16).

The first-principles calculation of the folding free-energy surface (also known as the potential of mean force because it represents the potential whose gradient is the mean force along the chosen reaction coordinate) for fragment B of protein A in an explicit solvent environment was performed as follows. We defined the reaction coordinate, onto which we projected the comput-

ed free energy and related dynamic and thermodynamic properties, as the radius of gyration (R_{gy}), because this coordinate is directly related to the volume of the protein during the folding process. Several independent molecular-dynamics (MD) simulations, starting from the folded NMR structure, were performed to generate a database of structures that spanned the reaction coordinate from the native state to the unfolded state. In portions of these simulations, performed for >9 ns, increased temperature or biasing potentials were used to enhance the range of conformations sampled (17). This database was divided, according to the R_{gy} value, into 20 equally spaced partitions from ~9.3 to 14 Å in R_{gy} . The resulting structures within each partition were clustered on the basis of a dissimilarity function that incorporates core side-chain packing, helical hydrogen bonding, and solvent-accessible surface area with hierarchical, agglomerative clustering (18) (Fig. 1, legend). For each partition, a "natural" clustering was extracted from the hierarchy with an information function (19). This process resulted in the identification of between two and seven clusters within each partition. The structures closest to the cluster centers were identified (78 in all) and used as initial conditions in "importance-sampling" MD simulations. A histogram in R_{gy} was generated from each simulation. The overlapping histograms spanned the reaction coordinate. Additionally, the continuity in configurations from adjacent initial conditions was examined to ensure local reversibility in the underlying sampling. The free-energy surface was constructed from these histograms by the weighted-histogram method (20).

From the clustering alone, several features of the conformational free-energy surface are apparent. Fragment B of protein A shows characteristics of folding by means of a "folding funnel," as shown by the probability distribution (P) of structural dissimilarity (D_s) for the protein, $P(D_s|R_{gy})$, versus the folding-unfolding reaction coordinate, R_{gy} (Fig. 1). The term "folding funnel" (8) refers to a hierarchical ensemble of (interconverting) protein conformations that descend in energy, and simultaneously in number, toward a unique compact state as the protein folds. The distribution shown in Fig. 1 is unimodal and narrow for values of R_{gy} of <10.75 Å. At R_{gy} values of >10.75 Å, the distribution becomes increasingly broad and uniform, reflecting the fact that fragment B of protein A can adopt an increasing continuum of different conformations as it unfolds. Thus, the ensemble of structures we observe, coupled with the free-energy surface described below, form a funnel-like hierarchy as a function of R_{gy} , and not a directed sequential pathway. If there were, instead, a highly directed se-

Department of Molecular Biology, Scripps Research Institute, La Jolla, CA 92037, USA.

*Present address: Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

†To whom correspondence should be addressed.

quential pathway from the folded to the unfolded manifold of states, one would expect a compact distribution around a single value of conformational dissimilarity, as observed in the folded manifold for $R_{gy} < 10.75 \text{ \AA}$, to occur throughout the range of the folding coordinate. Multiple sequential pathways would look much the same. Despite the lack of a directed pathway, we do observe that a sequence of consensus interactions involving progressively more tertiary contacts develops as folding progresses.

Four sets of representative cluster centers are shown in Fig. 2. Each structure within a set represents an ensemble of similar protein conformations. These sets typify the three thermodynamic states from our free-energy surface calculation: (i) an unfolded manifold with little structure; (ii) an intermediate region, in which significant amounts of secondary structure exist but the number and probability of tertiary interactions are low; and (iii) the native manifold.

In the unfolded manifold, volumes greater than 2.2 times that of the native manifold, the core of the protein is substantially disrupted and solvated (Fig. 2D). In each of the three helical regions, both helical and nonhelical conformations of the peptide are apparent. Calculations of the helix content, averaged over conformations sampled during the free-energy calculation, indicate that the structural fluctuations at these large volumes are four times the size of those at native volumes. In this least compact manifold of states, the formation of structure around the NH_2 -terminal reverse-turn is observed. Two proline-containing reverse-turn regions separate the three helices, those containing Pro^{21} and Pro^{39} , respectively. Of the two, the Pro^{21} region appears to be important in the early folding of this protein. Between R_{gy} values of 12 and 12.5 \AA , the NH_2 -terminal reverse-turn folds and brings helices I and II, which are partially folded, into contact in 84% of the clusters. At this stage in folding, Phe^{14} , Ile^{17} , Leu^{20} , Leu^{23} , Phe^{31} , Ile^{32} , and Leu^{35} participate in hydrophobic contacts and appear to be the reason that structure forms around the NH_2 -terminal turn.

As the overall structure becomes more compact, at R_{gy} values of ~ 11 to 12 \AA , helices I and II pack together into an antiparallel helix-turn-helix subdomain and the amount of secondary structure increases by 33% relative to the unfolded region. The interhelical contacts do not appear to be very specific, and the helices fluctuate in their relative orientation and registration at this stage of folding. Specific interhelical contacts involving the side chains noted above are explored for the packing of helices I and II. Helix III remains, for the most part, separate from the helix I-II subdomain (Fig. 2C).

At R_{gy} values of $< 11.0 \text{ \AA}$, the structures are highly compact but malleable in shape. The three-helix bundle topology is apparent. The helical content of helix III fluctuates by one or two turns, and helix II de-

velops a kink to allow for optimal core packing. Helix I is still associated with helix II, and it is in the final stages of folding that helix I changes its orientation relative to helix II and comes to lie across helix III,

Fig. 1. The distribution of structural dissimilarity between conformations with a common value of protein R_{gy} , $P(D_s|R_{gy})$. This distribution indicates that fragment B of protein A folds by descending a funnel-like series of conformational states. The distribution defines the range of different conformations (conformational dissimilarity) sampled by the protein as the simulations probe different regions of the folding-unfolding reaction coordinate, R_{gy} (measured in angstroms). The conformational dissimilarity function used to construct this distribution is defined as a dimensionless measure of structural difference:

$$D_s(A,B) = \sum \frac{(d_{ij}^A - d_{ij}^B)^2}{\text{Var}(d_{ij})} + \sum \frac{(d_i^A - d_i^B)^2}{\text{Var}(d_i)} + \frac{(\sqrt{SA^A} - \sqrt{SA^B})^2}{\text{Var}(SA)}$$

Thus, $D_s(A,B)$ is the distance between two structures A and B, d_{ij}^α is the interatomic contact distance between core side chains i and j in structure α , d_i^α is the hydrogen bond distance for the i th helical hydrogen bond in structure α , and SA^α is the solvent-accessible surface area for structure α .

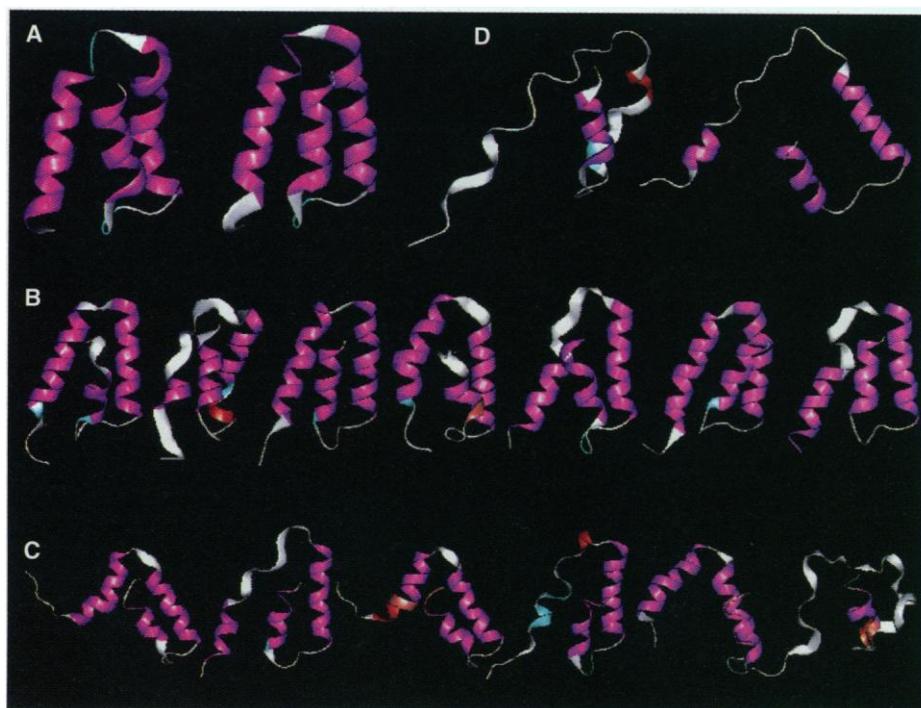
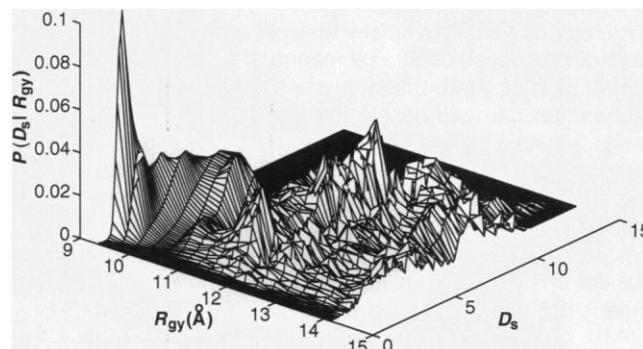


Fig. 2. The cluster centers representing four partitions of R_{gy} along the folding-unfolding coordinate. (A) Cluster centers for $R_{gy} = 9.37 \text{ \AA}$; (B) cluster centers for $R_{gy} = 10.13 \text{ \AA}$; (C) cluster centers for $R_{gy} = 12.13 \text{ \AA}$; and (D) cluster centers for $R_{gy} = 14.13 \text{ \AA}$. The orientation of structures in each panel place the NH_2 -terminal helix (helix I) near the center and the COOH -terminal helix (helix III) on the left. The colors in the figure indicate the presence of secondary structure types: magenta = α helix, red = 3_{10} helix, blue = turn, and white = coil. The cluster centers are from the database generated for initial conditions as described (17). The 20 equally spaced partitions in R_{gy} contained an average of ~ 350 structures each. Hierarchical clustering was carried out with the method of agglomerative nesting with Ward's distance function (19, 26). The natural clustering and corresponding cluster centers for each partition in R_{gy} were identified as described (18).

making a 15° to 20° crossing angle relative to the helix II axis (Fig. 2, A and B). Helix III adopts its final configuration relative to helices I and II by means of two mechanisms. Either helical hydrogen bonds develop from the NH₂-terminus of helix III, in which case the helix folds and then docks onto the subdomain in an approximately rigid-body fashion, or one of the core residues at the end of the unwound COOH-terminal strand makes contact with some of the core residues in the NH₂-terminal reverse-turn or base of helix I. After contact formation, the COOH-terminal strand folds into a helix and docks simultaneously. This process appears to be highly cooperative (Fig. 2B).

From these observations and detailed analysis of tertiary contact formation along the thermodynamic reaction coordinate used here, R_{gy} , we suggest that the early thermodynamic stages of folding should be resistant to H-D exchange in the NH₂-terminal region of the polypeptide chain, centered around Pro²¹, and the helix I-helix II interface. Significant resistance in

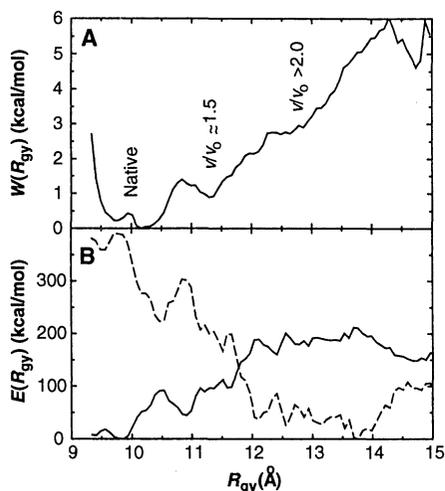


Fig. 3. (A) The free-energy surface, $W(R_{gy})$, and (B) two components [protein-protein (solid line) and protein-solvent (dashed line)] of the average energy, $E(R_{gy})$, displayed as a function of R_{gy} . These properties illustrate the dependence of folding thermodynamics on this (volume-related) reaction coordinate. The volume ratios v/v_0 indicate the volume of the protein (v) relative to that in the native state (v_0). The free-energy surface was constructed by the method of umbrella sampling with MD (27). For each of the 78 initial configurations of fragment B of protein A in explicit solvent, MD in the presence of a biasing potential was performed for a period of between 50 and 180 ps. The biasing potential, U_{bias} , is given by $U_{bias} = k(R_{gy} - R_{gy}^{ref})^2$, where R_{gy}^{ref} represents one of the R_{gy} partition centers at 9.38, 9.63, . . . 14.13 Å, and k is 5 to 15 kcal mol⁻¹ Å⁻². Histograms were collected in the R_{gy} coordinate. Energies and configurations of protein and solvent were saved every 10 steps to construct the average energy surfaces and compute the average properties described in the text.

helix III should occur only late in the thermodynamic folding pathway.

The free-energy and energy surfaces (Fig. 3) provide an energetic complement to the structural picture just described. The free energy as a function of R_{gy} shows three well-defined regions (Fig. 3A). There is a broad native-state basin. A barrier near $R_{gy} = 10.8$ Å, close to the native state, separates the native manifold from a small metastable minimum at ~ 1.5 times the native volume. The native state is ~ 1.1 kcal/mol lower in free energy than this intermediate state and is stable to equilibrium fluctuations on the order of $3k_B T$ at room temperature (k_B , Boltzmann constant; T , temperature). For values of R_{gy} greater than ~ 11.5 Å, the free energy increases by ~ 1.5 kcal/mol per angstrom.

The native state is favored by the protein-protein component of the energy and disfavored by the protein-solvent component from the energy surfaces for folding-unfolding (Fig. 3B). The protein-solvent energy dominates the energetic contributions to folding, rendering folding energetically unfavorable (Table 1) (see below). This observation is consistent with the structural features of unfolding: As the protein unfolds, it exchanges favorable protein-protein interactions with solvent; the protein becomes more exposed to solvent, thereby increasing the number of protein-solvent interactions and decreasing the overall protein-solvent energy. The core side chains are in contact with 55% more solvent in the unfolded state than in the native state, and all side chains show an increased exposure to solvent of $\sim 17\%$ relative to the native manifold. This increase in solvent exposure is accompanied by a decrease in the mobility of water around the protein side chains as the protein unfolds.

The overall thermodynamics for folding of fragment B of protein A are shown in Table 1. These properties are calculated to provide a qualitative molecular picture of folding-unfolding thermodynamics. The free-energy differences for the processes native \rightarrow intermediate and native \rightarrow unfolded indicate that the native state is ~ 1.1 kcal/mol more stable than the intermediate and 2.6 kcal/mol lower in free energy than the unfolded manifold. The energy and entropy of this system can each be decomposed into the sum of three terms: a solvent-solvent (SS) term, a protein-solvent (PS) term, and a protein-protein (PP) term. It can be shown that the solvent-solvent entropy and energy terms cancel exactly (21). The protein-solvent energy term represents the averaged energy of interaction between the protein and surrounding solvent, and favors more-unfolded states. The protein-protein energy term represents averaged intraprotein interactions and favors folding. On removal of the solvent-solvent terms, the unfolded state is favored energetically over the folded state by ~ 2.4 kcal/mol per residue. The corresponding entropy ($-T\Delta S_{NU}^{PP+PS}$) (Table 1) favors folding by approximately the same amount. The protein-protein entropy must increase as a function of the protein volume, disfavoring folding. This entropy term is qualitatively apparent in Fig. 1 as the widening of the distribution. The remaining protein-solvent entropy describes how the protein affects the solvent coordinate distribution and vice versa, the so-called hydrophobic effect (22). The overall magnitude of the entropy, which favors folding, has been suggested to be ~ 2 kcal/mol per residue of protein (23), in reasonable agreement with our findings. Also, general findings regarding the overall free energy of stability are consistent with the thermodynamic data for a synthetic, parallel three-helix bundle pro-

Table 1. Free-energy, energy, and entropy components for folding of fragment B of staphylococcal protein A. The folding free energy and component energies and entropies were computed from the underlying free-energy and energy surfaces shown in Fig. 3 (with the exception of the solvent-solvent energy surface for which the data are not shown). The Helmholtz free-energy difference (NVT ensemble) between the folded and intermediate (unfolded) manifolds of states was constructed from $\Delta A = -k_B T \ln K_{N(I,U)}^{eq}$ with $K_{N(I,U)}^{eq} = (\int_I \exp[-W(R_{gy})/k_B T] dR_{gy}) / (\int_U \exp[-W(R_{gy})/k_B T] dR_{gy})$; where the temperature (T) is 300 K, k_B is the Boltzmann constant, and the native manifold (N) was defined as $R_{gy} < 10.8$ Å, the intermediate manifold (I) as 10.8 Å $< R_{gy} < 12$ Å, and the unfolded manifold (U) as $R_{gy} > 12$ Å. The energy differences (ΔE) follow from averaging the energy surfaces over the corresponding regions of R_{gy} and the entropy (ΔS) comes from $\Delta A = \Delta E - T\Delta S$. The error estimates were constructed from the variance of batch means. The raw data used to compute $W(R_{gy})$ and the energy surfaces were divided into four batches. The variance of batch values about the overall average is indicative of the precision of the calculations. This quantity is given as \pm variance in the table. The corresponding solvent-solvent (SS) energy contributions are $\Delta E_{N \rightarrow I}^{SS} = 38 \pm 12$ kcal/mol and $\Delta E_{N \rightarrow U}^{SS} = 117 \pm 12$ kcal/mol; these terms are canceled in the overall thermodynamics of folding-unfolding by corresponding entropic terms. The thermodynamic driving forces for folding must arise from the terms included in the table. PP, protein-protein; PS, protein-solvent. Units are kilocalories per mole.

Transition	ΔA	ΔE^{PP+PS}	ΔE^{PP}	$-T\Delta S^{PP+PS}$
$N \rightarrow I$	1.1 ± 0.1	-44.0 ± 3.6	50.1 ± 4.4	45.0 ± 3.7
$N \rightarrow U$	2.6 ± 0.1	-115.4 ± 4.8	139.1 ± 5.2	118.2 ± 4.8

tein of comparable size (24). The accord that emerges from our calculations for basic thermodynamic parameters suggests that studies such as those presented here can be used for two purposes: (i) to provide a bridge to analytic mean-field and lattice-based theories, furnishing them with fundamental parameters for their model descriptions (4, 25), and (ii) to provide predictions and rationalizations of experimental observations.

REFERENCES AND NOTES

- O. B. Ptitsyn, R. H. Pain, G. V. Semisotnov, E. Zernovnik, O. I. Razgulyaev, *FEBS Lett.* **262**, 20 (1990); O. B. Ptitsyn, *J. Protein Chem.* **6**, 273 (1987); V. N. Uversky, *Biochemistry* **32**, 13288 (1993); _____ and O. B. Ptitsyn, *FEBS Lett.* **321**, 15 (1994); K. Kuwajima, *Proteins Struct. Funct. Genet.* **6**, 87 (1989); N. Taddei *et al.*, *Eur. J. Biochem.* **225**, 811 (1994).
- E. I. Shakhnovich and A. V. Finkelstein, *Biopolymers* **28**, 1667 (1989).
- A. Kolinski and J. Skolnick, *Proteins* **18**, 353 (1994); J. Skolnick and A. Kolinski, *J. Mol. Biol.* **221**, 499 (1991).
- J. D. Bryngelson, J. N. Onuchic, N. D. Socci, P. G. Wolynes, *Proteins Struct. Funct. Genet.* **21**, 167 (1995); K. A. Dill *et al.*, *Protein Sci.* **4**, 561 (1995); M. Karplus and A. Sali, *Curr. Opin. Struct. Biol.* **5**, 58 (1995).
- R. L. Baldwin, *Trends Biochem. Sci.* **14**, 291 (1989).
- D. L. Weaver and M. Karplus, *Protein Sci.* **3**, 650 (1994).
- K. A. Dill, *Biochemistry* **29**, 7133 (1990).
- P. E. Leopold, M. Montal, J. N. Onuchic, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 8721 (1992); P. G. Wolynes, J. N. Onuchic, D. Thirumalai, *Science* **267**, 1619 (1995).
- J. Skolnick, A. Kolinski, C. L. Brooks III, A. Godzik, A. Rey, *Curr. Biol.* **3**, 414 (1993); M. Veith, A. Kolinski, C. L. Brooks III, J. Skolnick, *J. Mol. Biol.* **237**, 361 (1994).
- C. L. Brooks III, *Curr. Opin. Struct. Biol.* **3**, 92 (1993); V. Daggett and M. Levitt, *ibid.* **4**, 291 (1994).
- C. L. Brooks III, *J. Mol. Biol.* **227**, 375 (1992).
- _____ and D. A. Case, *Chem. Rev.* **93**, 2487 (1993).
- Fragment B of staphylococcal protein A is a single-domain three-helix bundle protein that comprises 60 amino acids. The sequence of the 60-residue protein fragment is (key hydrophobic core residues are shown as bold) TADNKFNKEQQNAFYELH-LPNLNFEQRNGFIQSLKDDPSQSANLLAEAKKL-NDDAQAPKA. (Abbreviations for the amino acid residues are A, Ala; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; and Y, Tyr.) In solution (16), the COOH-terminal helix III is intact and all three helices contact one another, forming a small (~30 side-chain contacts) hydrophobic core. In complex with an immunoglobulin (15), helix III is basically unfolded, even though it is not in direct contact with the protein, and helices I and II pack together, resulting in a significantly smaller interhelical angle and differing contacts. Our simulations are of the sequence from residues 10 to 55 (underlined), which includes the three helices and begins at the NH₂-terminal Gln¹⁰ and ends at the COOH-terminal Ala⁵⁵. In the simulations, the NH₂-terminus and COOH-terminus of the polypeptide were blocked with acetyl and *N*-methyl amine groups, respectively.
- J. Sjodahl, *Eur. J. Biochem.* **78**, 471 (1977).
- J. Deisenhofer, *Biochemistry* **20**, 2361 (1981).
- H. Gouda *et al.*, *ibid.* **31**, 9665 (1992).
- To generate initial conditions spanning a relevant range of R_{gy} , six different MD simulations were performed. Each simulation consisted of a system defined by the protein plus 5012 water molecules in a periodic volume. The protein and solvent interacted via the CHARMM version 19 empirical force field [B. R. Brooks *et al.*, *J. Comput. Chem.* **4**, 187 (1983)], in which water is represented by the TIP3P model [W. L. Jorgensen, J. Chandrasekhar, J. Madura, M. L.

Klein, *J. Chem. Phys.* **79**, 926 (1983)]. SHAKE was used to maintain hydrogen-heavy atom bond distances fixed and to render the water molecules rigid at their experimental geometry [J.-P. Ryckaert, G. Cicotti, H. J. C. Berendsen, *J. Comput. Phys.* **23**, 327 (1977)]. Long-range forces were truncated at 10 Å with an atom-based force-shift method [C. L. Brooks III, B. M. Pettitt, M. Karplus, *J. Chem. Phys.* **83**, 5897 (1985)]. Interactions were computed with lists updated every 20 time steps [M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Oxford University Press, Oxford, 1989); L. Verlet, *Phys. Rev.* **159**, 98 (1967)], and the time step for integration of the equations of motion was 2.0 fs. Two independent 1.3-ns simulations were performed on the "native" state at 300 K, starting from the NMR structure. The temperature was smoothly increased to 360 K and maintained there for another 800 ps. Each system was then heated to 400 K and held there for 800 ps. The trajectories were analyzed to extract a conformational database that spanned the range of R_{gy} from 9.3 to 12.5 Å. Sampling in large-protein volume regions was accomplished with four additional simulations initiated from structures, with both large and small R_{gy} , samples in the previous simulations. These structures were unfolded slowly and stepwise with a harmonic potential in R_{gy} centered around a value of 14 Å. In each instance, a small initial force constant was chosen and was increased as the simulation continued. After the structures reached the target value of R_{gy} , the biasing potential was removed and the structures were allowed to relax. Each of these simulations required between 300 and 500 ps to achieve the target value. These structures were added to the database. The database initially contained ~300,000 structures. Conformations that were nearest neighbors in time were eliminated to diminish correlation and reduce the database to a manageable size.

18. To extract the "natural" clustering from the hierarchy, we used the function

$$E(m) = \frac{MBCD(m) - MBCD(m+1)}{\sqrt{SSQ(m)} - \sqrt{SSQ(m+1)}}$$

which peaks strongly for clusterings that have compact, well-separated clusters (19). In this expression, $MBCD(m)$ is the minimum between-cluster distance for clusters at level m of the hierarchy, and $SSQ(m)$ is the mean sum of squared deviation between all structures at the m th level of the hierarchy. These clusters, and their representative centers, define the range of conformations sampled during folding. These structures form the basis for our analysis of the folding funnel shown in Fig. 1 and the folding pathway described in Fig. 2.

- S. Xu, M. V. Kamath, D. W. Capson, *Patt. Recog. Lett.* **14**, 7 (1993); T. Kurita, *Patt. Recog.* **24**, 205 (1991).
- E. M. Boczkó and C. L. Brooks III, *J. Phys. Chem.* **97**, 4509 (1993); S. Kumar, J. Bouzida, R. Swendsen, P. Kollman, J. Rosenberg, *J. Comp. Chem.* **13**, 169 (1992).
- H.-A. Yu and M. Karplus, *J. Chem. Phys.* **89**, 2366 (1988).
- T. Lazaridis and M. E. Paulaitis, *J. Phys. Chem.* **96**, 3847 (1992).
- B. Honig, paper presented at the Cold Spring Harbor meeting on Protein Design/Folding, Cold Spring Harbor, NY, 18 October 1994; P. Alexander, S. Fahnestock, T. Lee, J. Orban, P. Bryan, *Biochemistry* **31**, 3597 (1992); S. J. Gill and P. L. Privalov, *Adv. Protein Chem.* **39**, 191 (1988).
- M. Lieberman, M. Tabet, T. Sasaki, *J. Am. Chem. Soc.* **116**, 5035 (1994).
- Z. Luthey-Schulten, R. E. Ramirez, P. G. Wolynes, *J. Phys. Chem.* **99**, 2177 (1995).
- A. D. Gordon, *J. R. Stat. Soc. A* **150**, 119 (1987).
- J. P. Valleau and G. M. Torrie, in *A Guide to Monte Carlo for Statistical Mechanics: 2. Byways in Statistical Mechanics, Part A*, B. J. Berne, Ed. (Plenum, New York, 1977), pp. 169–194.
- We thank the Pittsburgh Supercomputing Center and the NSF Meta-Center Allocations Board for allocations of computing time. Supported by NIH grant GM48807.

3 January 1995; accepted 30 May 1995

Telomerase in Yeast

Marita Cohn and Elizabeth H. Blackburn*

The ribonucleoprotein enzyme telomerase synthesizes telomeric DNA by copying an internal RNA template sequence. The telomerase activities of the yeasts *Saccharomyces castellii* and *Saccharomyces cerevisiae*—with regular and irregular telomeric sequences, respectively—have now been identified and characterized. The *S. cerevisiae* activity required the telomerase RNA gene *TLC1* but not the *EST1* gene, both of which are required for normal telomere maintenance in vivo. This activity exhibited low processivity and produced no regularly repeated products. An inherently high stalling frequency of the *S. cerevisiae* telomerase may account for its in vitro properties and for the irregular telomeric sequences of this yeast.

Telomeres, the specialized DNA-protein structures at the ends of eukaryotic chromosomes, are necessary for chromosomal stability (1). The telomeric DNA that is essential for telomere function consists of simple sequences, repeated in tandem, whose synthesis and maintenance normally require the ribonucleoprotein enzyme te-

lomerase (2–5). Telomerase activities identified in vitro in extracts from ciliated protozoan and vertebrate species catalyze the synthesis of homogeneous, precisely repeated telomeric sequences (6, 7). Synthesis by telomerase occurs by addition of the G-rich telomeric strand to the 3' end of a DNA primer. For the *Tetrahymena thermophila* telomerase, both in vitro and in vivo studies show that DNA synthesis occurs by copying a short template sequence within the RNA moiety of the enzyme (2, 8).

Department of Microbiology and Immunology, University of California, San Francisco, CA 94143–0414, USA.

*To whom correspondence should be addressed.