M. Vanetti, *et al.*, *FEBS Lett.* **311**, 290 (1992); K. Yasuda *et al.*, *J. Biol. Chem.* **267**, 20422 (1992); W. Meyerhof, *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 10267 (1992); J. F. Bruno, Y. Xu, J. Song, M. Berelowitz, *ibid.*, p. 11151; L. Rohrer *et al.*, *ibid.* **90**, 4196 (1993); A. M. O'Carroll, *et al.*, *Mol. Pharmacol.* **42**, 939 (1992).
8. B. Hunyady, *et al.*, *Endocrinology* **138**, 2632 (1997).
9. S. W. Mitra *et al.*, *ibid.*, in press.
10. E. Mezey *et al.*, *ibid.* **139**, 414 (1998).
11. L. Yang *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 10836 (1998).
12. K. Raynor *et al.*, *Mol. Pharmacol.* **43**, 838 (1993).
13. W. J. Rossowski and D. H. Coy, *Biochem. Biophys. Res. Commun.* **205**, 341 (1994).
14. L. Yang *et al.*, *J. Med. Chem.* **41**, 2175 (1998).
15. Coupling of sst1 to adenylate cyclase was not observed with our hsst1 CHO-K1 cell line. Our result is consistent with results obtained by S. Rens-Domiano *et al.* [*Mol. Pharmacol.* **42**, 28 (1992)] and K. Raynor *et al.* (*12*). In contrast, Y. C. Patel, M. Greenwood, A. Warszynska, R. Panetta, and C. B. Srikant [*Biochem. Biophys. Res. Comm.* **198**, 605 (1994)] have observed coupling of the sst1 receptor to adenylate cyclase in CHO cells.
16. L cells are tissue culture cells originally derived from murine connective tissue. They are fibroblast-like in their morphology. The cell line was developed in the mid-1940s. The letter L is a strain designation assigned by the developers of the line.
17. Y. C. Cheng and W. H. Prusoff, *Biochem. Pharmacol.* **22**, 3099 (1973).
18. CHO-K1 cells stably expressing the sst2, sst3, sst4, and sst5 receptors were subcultured in 12-well plates and treated at 37°C for 30 min with growth medium containing 0.5 mM isobutylmethylxanthine. The medium was replaced with fresh growth medium, with or without 10 μM forskolin and test agents. After incubation for 5 min at 37°C, the medium was removed, and the cells were lysed by freeze-thaw in 0.1 M HCl. The cAMP content of duplicate wells was determined with a radioimmunoassay kit (Amersham). Data obtained from the dose-response curves were analyzed by nonlinear regression with GraphPad Prism, version 2.01 (GraphPad Software, San Diego, CA).
19. We thank K. Cheng, L.-Y. Pai, and T.-J. Wu for growth hormone secretion assays.

19 June 1998; accepted 22 September 1998

Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution

Yong Duan and Peter A. Kollman*

An implementation of classical molecular dynamics on parallel computers of increased efficiency has enabled a simulation of protein folding with explicit representation of water for 1 microsecond, about two orders of magnitude longer than the longest simulation of a protein in water reported to date. Starting with an unfolded state of villin headpiece subdomain, hydrophobic collapse and helix formation occur in an initial phase, followed by conformational readjustments. A marginally stable state, which has a lifetime of about 150 nanoseconds, a favorable solvation free energy, and shows significant resemblance to the native structure, is observed; two pathways to this state have been found.

Elucidation of the mechanism of protein folding is an important step in understanding the relation between sequence and structure of proteins. Understanding of the mechanism should allow more accurate prediction of protein structures, with wide-ranging implications in biochemistry, genetics, and pharmaceutical chemistry. The recent hypothesis of folding-related diseases is another example of the significance of folding (*1*). Yet despite great progress made by a variety of experimental and theoretical studies after decades of extensive research, it has been difficult to establish detailed descriptions of the folding process and mechanism (*2*).

Computer simulation of molecular systems can provide rich information at various levels of resolution, and this approach has been important in attempts to understand protein folding mechanisms. A simplified representation might treat the protein residues as (one or two) linked beads (*3*). Higher resolution models represent most or all of the atoms of the protein explicitly, with an implicit representation of the solvent (*4*, *5*). At an even higher level of detail are molecular dynamics (MD) simulations with full atomic representation of both protein and solvent. Such calculations are uniquely suited to the study of protein folding because of their resolution and accuracy. The simulation parameters (that is, the force field) are derived from experiments and from gas-phase quantum mechanical calculations and have been tested in smaller systems in many critical comparisons with experimental results (*6*). Because of the complexity of the representation, the large number of atoms (often exceeding 10,000), and the need to take time steps of 1 to 2 fs, such simulations have, to date, been limited to a few nanoseconds (*7*) (or a few million integration steps). This has precluded the simulation of even the early stages of protein folding. Nevertheless, insights have been gained from unfolding simulations (*8*) of the denaturation process. Attempts have also been made to construct the folding free-energy landscape from unfolding simulations (*9*). Direct folding simulations with this approach, however, have been limited to small peptide fragments and have been carried out for as long as 50 ns (*10*). Direct simulation of the protein folding process with such an approach has not been considered possible (*11*) "either now or in the foreseeable future" (*12*, p. 29). By using a Cray T3E, a massively parallel supercomputer consisting of hundreds of central processing units (CPUs) connected by low-latency, high-speed, and high-availability networks, with an efficiently parallelized program that scales well to the 256-CPU level for small protein–solvent systems and is six times faster than a typical current state-of-the-art program (*13*), we have conducted a 1-μs simulation at 300 K on the villin headpiece subdomain, a 36-residue peptide (HP-36) (*14*, *15*), starting from a fully unfolded extended state (Fig. 1A), including ~3000 water molecules (*16*–*18*). The simulation time scale is close to that required to fold small proteins. The simulation shows a mechanism for the protein to find a folding intermediate, an important step in proceeding to its fully folded state.

Proteins can have marginally stable nonnative states that are difficult to observe experimentally (*19*). Computer simulation can play an important role in identifying these structures because of its extremely high time resolution and detailed atomic level representation. Recent experimental studies suggest that the time for (small) proteins to reach their marginally stable states with partially formed secondary structures is on the order of 10 μs (*20*). It also has been shown that a small protein can fold within 20 μs (*21*), and it has been estimated that the lower limit of the folding time is 1 μs (*22*).

HP-36 is one of the smallest proteins that can fold autonomously. It contains only naturally occurring amino acids and does not require disulfide bonds, oligomerization, or ligand binding for stabilization; its melting temperature is above 70°C in aqueous solution (*14*). The estimated folding time of the protein is between 10 and 100 μs (*23*), which would make it one of the fastest folding proteins. Nuclear magnetic resonance (NMR) studies of the 36-residue subdomain revealed three short helices (Fig. 1C) (*15*). We refer to them as helices 1, 2, and 3, for residues 4 to 8, 15 to 18, and 23 to 30, respectively, as found in the NMR structure. They are held together by a loop (residues 9 to 14), a turn (residues 19 to 22), and

Department of Pharmaceutical Chemistry, University of California, San Francisco, CA 94143, USA.

*To whom correspondence should be addressed. E-mail: pak@cgl.ucsf.edu

a closely packed hydrophobic core. In the following, we number the residues from 1 to 36, where our residue 1 corresponds to residue 41 in the NMR structure. The unfolded starting structure (Fig. 1A), which was generated from the NMR native structure by a 1.0-ns MD simulation at 1000 K, was in an extended state with very few native contacts (<3%) and no helical content.

In addition to the 1-μs simulation, a control simulation was conducted for 100 ns at 300 K (*16*), starting from the native NMR structure (*15*). In this 100-ns simulation, the NH_2-terminal helix 1 rotated ~30° outward while maintaining its helical structure. The COOH-terminal residue Phe^{36}, which was disordered in the NMR structure, also exhibited large-scale movement. Phe^{36} was initially in the solvent, as found in the NMR structure. Together with Leu^{35}, it soon moved toward the COOH-terminus of helix 1 and loosely packed against the middle of Lys^{8}, forming a small hydrophobic cluster comprising Lys^{8}, Leu^{35}, and Phe^{36}. Judging from the reduction of the hydrophobic surface, the formation of these contacts appears energetically reasonable. The overall structure, particularly the middle portion (helices 2 and 3) and the hydrophobic core, remained stable in the simulation. The average root mean square deviation (rmsd) from the NMR structure was 1.5 Å for the main chain atoms of residues 9 to 32 in the last 50 ns of the trajectory, whereas this rmsd varied from 3.0 to 8.8 Å during the last 800 ns of our 1-μs folding trajectory. The fact that the core of the native structure remained near the NMR structure indicated that our simulation protocol was adequate to study protein folding.

In the 1-μs trajectory, the radius of gyration (R_g) fluctuated (Fig. 2C) between 16 Å, which represents extended states, and 8.7 Å, which represents highly compact states, compared with 9.4 Å of the native structure. The main-chain rmsd (Fig. 2C) of all residues (1 to 36) varied between 12.4 and 4.5 Å; that of the middle portion (residues 9 to 32) fluctuated between 8.8 and 3.0 Å. Up to 80% of the native helical content (Fig. 2A) and up to 62% of the native contacts (Fig. 2B) were formed. The solvation component of the free energy (SFE) (Fig. 2D) also reached levels comparable to that of the native structure. More importantly, a marginally stable state was reached, as can be seen from the rmsd's and the R_g, which had a residence time of longer than 150 ns, much longer than typical MD simulations conducted to date.

An important feature of most of the trajectory is its high degree of fluctuation, exhibited by essentially all the features measured, including native helical content, native contacts, rmsd, and R_g (Fig. 2). Such a large degree of fluctuation is in contrast to the relatively small fluctuations found during the simulation beginning from the native structure and during the time when the marginally stable state was reached in the folding simulation. This high degree of fluctuation is an indication of the rugged and shallow free-energy landscape associated with early stages of folding. This shallow landscape enables the protein to search the early-stage folding free-energy surface easily.

The folding began with a "burst" phase, characterized by a steady rise in native helical content (Fig. 2A) and in native contacts (Fig. 2B), and the decrease of the SFE (Fig. 2D), which lasted from the beginning of the trajectory to ~60 ns. Within this period of time, the native helical content increased to ~60% from an initial value of zero; meanwhile, the native contacts increased to about 45% from an initial value of 3%, and the SFE was reduced by nearly 14 kcal/mol, reaching a level comparable to that of the native structure. Analysis of the correlations between various energy terms and R_g indicated that the initial phase of the 300 K simulation was driven by the burial of exposed hydrophobic surface (*13*). Therefore, this phase can be seen as an initial hydrophobic collapse. However, given the concomitant rise of the helical content, it appears that hydrophobic collapse occurs on the same time scale as formation of some secondary structure. This makes physical sense in that a protein, as it buries its hydrophobic groups, tries to avoid burying its hydrogen bonding functionalities, and secondary structure formation provides a way to do this. The time required to reach 50% helical content was about 60 ns, in excellent agreement with recent kinetic measurements on apo-myoglobin (48 ns) (*24*) and on alanine peptide (16 to 180 ns) (*25*). Given the diverse folding rates observed in different sequences in experiments (16 ns for alanine peptide and 48 ns for apo-myoglobin, with the same method), the small difference observed here may be the result of sequence dependence.

Alonso and Daggett (*26*) showed that almost all nonnative conformations of ubiquitin generated in unfolding simulations moved to a lower R_g when the temperature was lowered. Their least native structure went through two cycles of expansion and collapse in 2 ns. Our simulations expand on these findings by showing that cycles of expansion and collapse can extend even into the microsecond regime and that these expanded and collapsed structures get more nativelike as folding proceeds.

This burst phase was followed by a slower adjustment phase. The slower phase started from a sharp drop of the helical content, from an average of more than 50% down to about

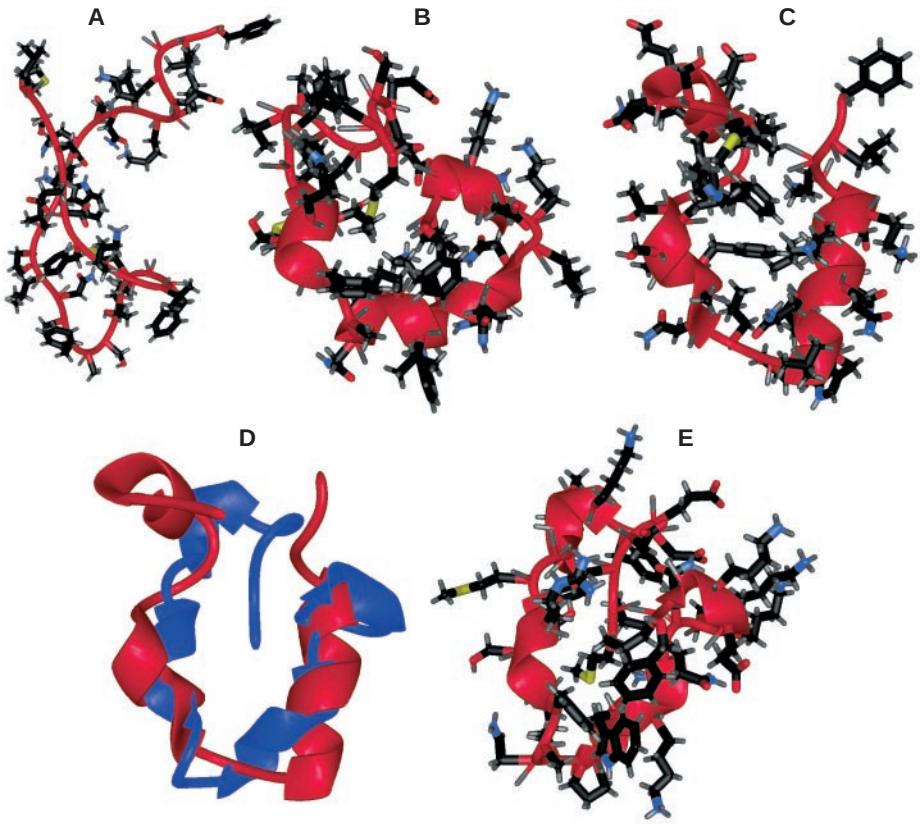

Fig. 1. Ribbon representations of (**A**) the unfolded, (**B**) partially folded (at 980 ns), and (**C**) native structures, and (**E**) a representative structure of the most stable cluster and (**D**) the overlap of the native (red) and the most stable cluster (blue) structures, generated with UCSF MidasPlus. Color code [except (D)]: red, main chain atoms and oxygen; black, non–main chain carbon; blue, non–main chain nitrogen; gray, hydrogen; yellow, sulfur.

20%, while the R_g decreased slightly and the SFE became more favorable. Meanwhile, the native contacts remained at a level similar to the one at the end of the burst phase. After this initial drop of the helical content, both the helical content and the native contacts remained at their respective levels. They showed a steady but slow rise after about 200 ns. A two-phase folding process has also been observed in the kinetic measurement of apo-myoglobin in which a fast 48-ns burst phase was followed by a 132-μs slow phase that was interpreted as the tertiary-contacts formation phase (*24*).

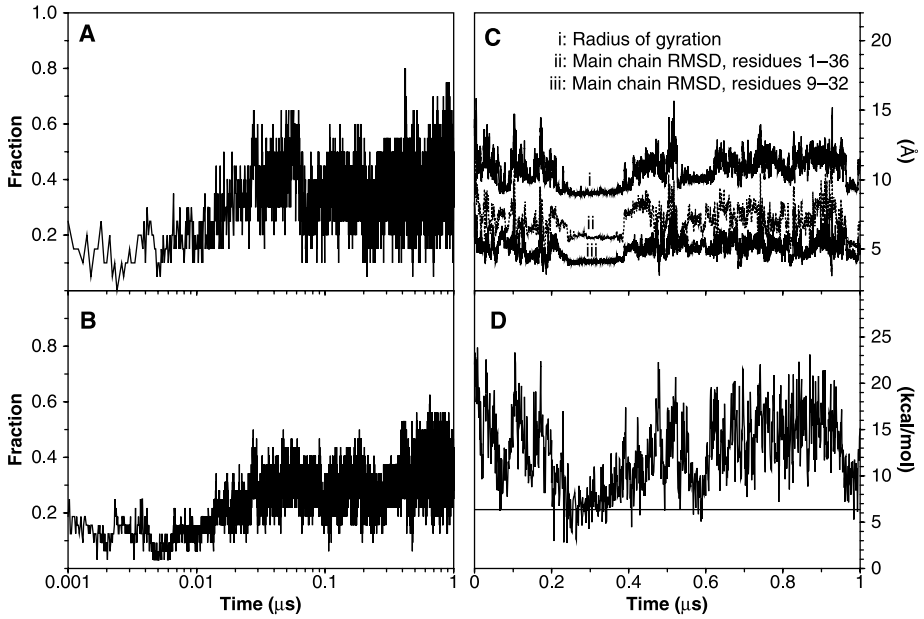

Fig. 2. Time evolution of (**A**) fractional native helical content, (**B**) fractional native contacts, (**C**) R_g and the main chain rmsd from the native structure, and (**D**) SFE of the protein. The helical content and the native contacts are plotted on a logarithmic time scale. The helical content was measured by the main chain ϕ-ψ angle ($-60° \pm 30°$, $-40° \pm 30°$). The native contacts were measured as the number of neighboring residues present in 80% of the last 50 ns of the native simulation. Residues are taken to be in contact if any of the atom pairs are closer than 2.8 Å, excluding residues i and $i+1$, which always have the contacts through main chain atoms. The SFE was calculated as described by Eisenberg and McLachlan (*31*) using their parameters (0.0163, −0.00637, 0.02114, −0.02376, and −0.05041, in kcal mol Å^{-2}, for the surface areas of nonpolar, polar, sulfur, charged oxygen, and charged nitrogen, respectively). The straight line represents the SFE of the native structure.

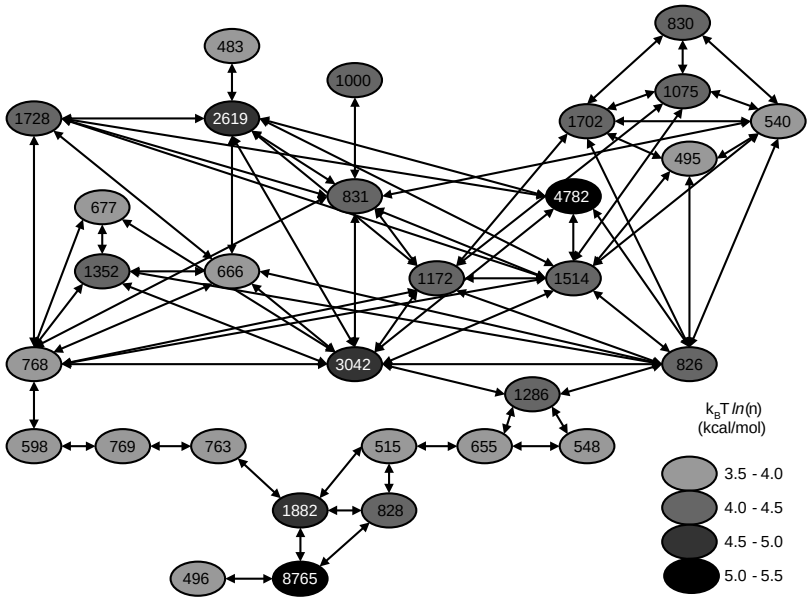

Fig. 3. Pathways of folding events. Circles represent the most populous clusters and arrows represent transitions between them. The circles are shaded by $k_BT \ln(n)$, where n is the number of snapshots in the cluster (see the legend), T is the temperature, and k_B is the Boltzmann constant. The label in each circle indicates the number of snapshots in the cluster.

To reach their folded states from fully unfolded states, proteins can (*27*) sample marginally stable states, which can be identified by clustering methods (*28*). The population of snapshots in each cluster reflects the likelihood of each cluster being sampled in the duration of the simulation. There are 13 clusters that each contain more than 1000 snapshots (or 2% of the trajectory, equivalent to 20 ns). Among these 13 most populous clusters, 10 are compact, with values of R_g between 9.1 to 10.4 Å (or 96 to 110% of the native R_g). A common feature of these clusters is the formation of a helix at the NH_2-terminus of helix 3, residues 23 to 28. Helix 2 is also partially formed in 12 of the 13 most populous clusters. This suggests that helix 2 and the NH_2-terminus of helix 3 are the initiation sites of folding.

The core regions of most of these 13 populous clusters are packed and inaccessible to water. Their average buried surface areas were more than 50% for residues 10 to 30, which make up the core region, compared with 58% burial of the same residues in the native structure. Only 4 of the 13 clusters showed more than 50% accessible surface for the same region. This is in contrast to the initial 200 ns of the trajectory, in which half of the compact structures showed a solvated core (*13*), indicating a shift toward compact structures that are less accessible to water as the folding progresses.

The most populous cluster had a population of 8765 snapshots, or 17.5% of the trajectory. Most of the snapshots in this cluster are found between 240 to 400 ns, while the rmsd and R_g are very stable, and the SFE is comparable to that of the native structure. It is also the most compact cluster whose R_g of 9.1 Å is smaller than that of the native structure (9.4 Å). Representative structures from this cluster show a marked similarity to the native structure (Fig. 1D). Among the secondary structure elements, helix 2 is well formed, helix 1 and 3 are partially formed, and the loop connecting helix 1 and 2 starts to form. The main chain rmsd of the structure shown in Fig. 1E relative to the NMR structure is 5.7 Å for all residues and 4.0 Å for residues 9 to 32. The SFE of this cluster is 7.8 ± 2.3 kcal/mol, close to the 7.1 ± 2.2 kcal/mol of the 100-ns native simulation trajectory and the 6.4 kcal/mol found for the experimental NMR structure. A notable feature of this cluster is its high stability. The longest residence time in the cluster is about 150 ns, much longer than that of any other state (typically a few nanoseconds).

While the simulation is in the most highly populated, nativelike state, the side chains do not reach their native positions (Fig. 1E). This is not surprising. It is unrealistic to expect the protein to fold to the native structure within our

simulation time of 1 μs, which is still much shorter than the lower bound of the estimated folding time, 10 μs (*23*). On the other hand, the estimated folding time is consistent with the formation of a marginally stable state that contains many of the features of the native structure. Because the residence time of the marginally stable state is longer than 150 ns, only two orders of magnitude shorter than the lower limit of the estimated folding time, and is highly nativelike, it may well be an intermediate state. We speculate that a number of intermediates such as the one we have observed will form and dissipate, until one forms that allows the precise side chain packing that will lead to the native state.

These states, identified as clusters, when linked together by the transitions, show the pathways of the folding events, whereas the number of transitions between clusters indicates the likelihood of such transitions in the simulation time scale. We have identified the most populated clusters and the transitions between them (Fig. 3). These transitions are bidirectional (that is, the number of forward and backward transitions are similar). A noteworthy feature is that the access to the marginally stable state (discussed above) is limited, and only two primary pathways to it were observed in the simulation. This is in contrast to other states with similar (but smaller) populations that are much more accessible. For example, the second most populated state can be readily accessed through five pathways. Consequentially, such states are kinetically less stable and have a much shorter residence time (a few nanoseconds) than the marginally stable state, even though on the basis of our limited sampling they are only slightly less favorable thermodynamically [$-k_BT \ln(4782/8765)$, or about 0.4 kcal/mol]. We speculate that limited access to the folded state [such that folding takes place by way of a few pathways or a dominant pathway (*5*)] may serve to provide kinetic stability to the thermodynamically stable folded state. More importantly, through the transition network, early states can readily transit between one another, resulting in thoroughly tangled multiple pathways. Therefore, the emergence of such pathways may be a key feature of the funnel-shaped folding landscape, with the role of the folding intermediate being to merge the multiple pathways.

Among the native contacts, nine were in contact for more than 50% of the simulation time; they were "local" contacts (that is, less than four residues apart along the chain). Ten of the native contacts were formed for less than 10% of the simulation time, and seven of these poorly formed contacts were tertiary contacts (that is, more than five residues apart). Our simulation indicates that the tertiary contacts are less likely to form and be maintained in the early stages of folding. Therefore, the formation of tertiary contacts is likely to be the bottleneck of the folding process. These results are consistent with kinetic measurements of Plaxco *et al*. who found that the folding speed is primarily determined by the contact order (*29*)—the more "nonlocal" contacts a protein has, the slower it folds—suggesting the contribution of chain entropy loss to the free-energy barrier of folding.

Our results show that microsecond-scale simulations of small proteins in a fully solvated environment can be used to probe the early stages of the folding process. Although we have presented only a single trajectory here whose statistical significance cannot be assessed, we have carried out a second trajectory on HP-36 and one on protein G starting from unfolded states for ~100 ns (*30*). The nature of the burst phase (hydrophobic collapse accompanied by secondary structure formation) was similar to that reported here. In addition, even though the repeated increases and decreases in the R_γ during the 1-μs trajectory do not represent statistically independent events, they can be viewed as steps in the process of the protein finding its way from the fully unfolded to its fully folded state. With the further development of massively parallel supercomputers and constant improvement of the simulation methods, simulation times may be extended to cover the entire folding process of small proteins (tens of microseconds) within a few years. Equally exciting are the methods that have allowed experimentalists to study protein folding on a submicrosecond time scale. Thus, direct and realistic comparisons between experimental and simulation studies of protein folding may soon be made on the same time scale. These comparisons should provide a useful and critical assessment of the accuracy of the simulation models such as force fields and boundary conditions, and yield a microscopic understanding of the folding process valuable to theoreticians and experimentalists alike.

References and Notes

1. R. N. Sifers, *Nature Struct. Biol.* **2,** 355 (1995); S. B. Prusiner, *Science* **278,** 245 (1997).
2. Reviewed in K. A. Dill and H. S. Chan, *Nature Struct. Biol*. **4,** 10 (1997); C. M. Dobson and O. B. Ptitsyn, *Curr. Opin. Struct. Biol*. **7,** 1 (1997); E. Shakhnovich and A. R. Fersht, *ibid.* **8,** 65 (1998).
3. J. Skolnick and A. Kolinski, *Science* **250,** 1121 (1990); A. Sali, E. Shakhnovich, M. Karplus, *Nature* **369,** 248 (1994); K. A. Dill *et al*., *Protein Sci.* **4,** 561 (1995).
4. C. A. Schiffer, V. Dötsch, K. Wüthrich, W. F. van Gunsteren, *Biochemistry* **34**, 15057 (1995).
5. T. Lazaridis and M. Karplus, *Science* **278**, 1928 (1997).
6. T. Fox and P. A. Kollman, *Proteins* **25,** 315 (1996); P. A. Kollman, *Acc. Chem. Res.* **29,** 461 (1996).
7. The longest single MD trajectories of proteins with explicit water have been 5.4 ns [A. Li and V. Daggett, *Protein Eng.* **8,** 1117 (1995)].
8. V. Daggett and M. Levitt, *Proc. Natl. Acad. Sci. U.S.A.* **89,** 5142 (1992); J. Tirado-Rives and W. L. Jorgensen, *Biochemistry* **32,** 4175 (1993); V. Daggett and M. Levitt, *Curr. Opin. Struct. Biol*. **4,** 291 (1994).
9. Brooks and co-workers have attempted to reconstruct the folding free-energy landscape [E. M. Boczko and C. L. Brooks III, *Science* **269,** 393 (1995); F. B. Sheinerman and C. L. Brooks III, *J. Mol. Biol*. **278,** 439 (1998)] from the restrained unfolding simulations using the WHAM method [S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, J. M. Rosenberg, *J. Comp. Chem.* **13,** 1011 (1992); E. M. Boczko and C. L. Brooks III, *J. Phys. Chem.* **97,** 4509 (1993)].
10. E. Demchuk, D. Bashford, D. A. Case, *Fold. Des.* **2,** 35 (1997); X. Daura, B. Jaun, D. Seebach, W. F. van Gunsteren, A. E. Mark, *J. Mol. Biol*. **280,** 925 (1998).
11. F. B. Sheinerman and C. L. Brooks, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 1562 (1998).
12. E. I. Shakhnovich, *Curr. Opin. Struct. Biol.* **7**, 29 (1997).
13. Y. Duan, L. Wang, P. A. Kollman, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 9897 (1998).
14. C. J. McKnight, D. S. Doering, P. T. Matsudarira, P. S. Kim, *J. Mol. Biol*. **260**, 126 (1996).
15. C. J. McKnight, P. T. Matsudaira, P. S. Kim, *Nature Struct. Biol*. **4**, 180 (1997).
16. The force field [W. D. Cornell *et al*., *J. Am. Chem. Soc.* **117,** 5179 (1995)] of Cornell *et al*. was used with full representation of solvent with the TIP3P water model [W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, *J. Chem. Phys.* **79,** 926 (1983)]. Periodic boundary conditions were imposed by a nearest image convention in a truncated octahedron box. An 8 Å residue-based cutoff was applied to the long-range nonbonded protein-water and water-water interactions (both electrostatic and van der Waals). The intramolecular nonbonded interactions of protein were calculated without truncation. When applicable, temperature and pressure controls were imposed through use of Berendsen's algorithms [H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, J. R. Haak, *J. Comp. Phys.* **81,** 3684 (1984)]. The solute and solvent were separately coupled to a temperature bath with coupling constants of 0.1 ps. The pressure coupling constant was 20 ps. The trajectories were produced by numerical integration with the Verlet-*leapfrog* algorithm [L. Verlet, *Phys. Rev.* **159,** 98 (1967); C. L. Brooks III, M. Karplus, B. M. Pettitt, *Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics* (Wiley, New York, 1989)] by using a 2-fs time step. Bond constraints were imposed on all bonds involving hydrogen atoms with SHAKE [J.-P. Ryckaert, G. Ciccotti, H. J. C. Berendsen, *J. Comp. Phys.* **23,** 327 (1977)] and SETTLE [S. Miyamoto and P. A. Kollman, *J. Comp. Chem.* **13,** 952 (1992)].
17. Preparation: The starting coordinates were the NMR structure of villin headpiece subdomain by McKnight *et al.* [C. J. McKnight, P. T. Matsudaira, P. S. Kim, *Nature Struct. Biol*. **4,** 180 (1997)] {Protein Data Bank [F. C. Bernstein *et al*., *J. Mol. Biol.* **112,** 535 (1977)] access code 1vii}. It was denatured by carrying out a 1-ns simulation in water at 1000 K using constant volume. The denatured molecule was then immersed in a truncated octahedron water box constructed from a cubic box of 76.5 Å. A total of 6510 water molecules were retained. The excess water molecules were removed after ~20 ns when a semistable compact structure was formed, to reduce the computational cost. About 3000 water molecules were retained for the remainder of the simulation. Production: The simulation was started from an equilibration phase of 1.0 ns at 200 K and 1 atm pressure. The long equilibration phase was intended to mimic an equilibrated fully denatured state and for adequate solvation of the molecule and to minimize any instability caused by the high-temperature origin of the starting structure. The density of the system was initially 0.90 g/cm^3, increased to 1.05 g/cm^3 within 10 ps, and remained so for the remainder of the 1-ns solvent equilibration trajectory. The simulation was then conducted for 1.0 μs (0.5 billion integration steps). Temperature and pressure were controlled at physiological conditions (that is, 300 K and 1 atm) by the methods described above. Both temperature and density stabilized within 10 ps. The trajectory was saved at 20-ps intervals for the analysis. The entire simulation took ~2 months' CPU time on a 256-CPU Cray T3D and an equal amount of CPU time on a 256-CPU Cray T3E-600.
18. The use of nonbonded cutoffs in the simulation, as performed here, is an inherent source of "noise" in the simulation trajectory. Such noise is not as dele-

terious in small proteins with limited formal charges as it is in, for example, RNA and DNA [T. E. Cheatham III, J. L. Miller, T. Fox, T. A. Darden, P. A. Kollman, *J. Am. Chem. Soc.* **117,** 4193 (1995)], as demonstrated by the stability of the 100-ns simulation at 300 K started from the native NMR structure. But accurate inclusion of long-range electrostatic effects [U. Essmann *et al.*, *J. Chem. Phys.* **103,** 8577 (1995)] does provide an additional challenge for achieving high parallelism in the MD code. Work is in progress to include long-range electrostatic effects while still achieving a level of parallelism and speed comparable to that presented here.

19. O. B. Ptitsyn, *Curr. Opin. Struct. Biol.* **5**, 74 (1995).
20. R. M. Ballew, J. Sabelko, M. Gruebele, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 5759 (1996).
21. R. E. Burton, G. S. Huang, M. A. Daugherty, T. L. Calderone, T. G. Oas, *Nature Struct. Biol.* **4**, 305 (1997).
22. S. J. Hagen, J. Hofrichter, A. Szabo, W. A. Eaton, *Proc. Natl. Acad. Sci. U.S.A.* **93,** 11615 (1996); W. A. Eaton, V. Muñoz, P. A. Thompson, C. K. Chan, J. Hofrichter, *Curr. Opin. Struct. Biol.* **7,** 10 (1997).
23. K. W. Plaxco, personal communication (1998).
24. R. Gilmanshin, S. Williams, R. H. Callender, W. H. Woodruff, R. B. Dyer, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 3709 (1997).
25. S. Williams *et al.*, *Biochemistry* **35,** 691 (1996); P. A. Thompson, W. A. Eaton, J. Hofrichter, *ibid.* **36,** 9200 (1997).
26. D. O. V. Alonso and V. Daggett, *J. Mol. Biol.* **247,** 501 (1995); D. O. V. Alonso and V. Daggett, *Protein Sci.* **7,** 860 (1998).
27. C. B. Anfinsen, *Science* **181**, 223 (1973).
28. A total of 50,000 sets of coordinates were accumulated every 20 ps. They were clustered by comparison with the average coordinate of existing clusters using a 3.0 Å main-chain rmsd cutoff, similar to the method described by Karpen *et al.* [M. E. Karpen, D. J. Tobias, C. L. Brooks III, *Biochemistry* **32,** 412 (1993)]. Those that are within 3.0 Å main-chain rmsd from the average coordinates of the cluster are assigned to the cluster. A total of 98 clusters were produced. Thirty clusters were highly populated with ~500 or more coordinate sets and 13 clusters had more than 1000 sets of coordinates.
29. K. W. Plaxco, K. T. Simons, D. Baker, *J. Mol. Biol.* **277**, 985 (1998).
30. Y. Duan and P. A. Kollman, unpublished data .
31. D. Eisenberg and A. D. McLachlan, *Nature* **319**, 199 (1986).
32. Supercomputing time was provided by Cray Research, a subsidiary of Silicon Graphics, Inc. (SGI), and by the Pittsburgh Supercomputing Center (PSC). We are grateful to R. Roskies and M. Levine (PSC), J. Carpenter and H. Pritchard (SGI), and J. Wendoloski (AMGEN) for their support. We thank K. Dill, D. Agard, I. Kuntz, J. Pitera, and T. Cheatham for critical reading of the manuscript; L. Wang, C. Simmerling, M. Crowley, J. Wang, and W. Wang for stimulating discussions; and L. Chiche for the solvation free-energy calculation program. Graphics were provided by Computer Graphics Lab of the University of California, San Francisco (T. Ferrin, Principal Investigator, grant RR-1081). This work was supported in part by NIH grant GM-29072, by a University of California Biotechnology Star grant, and by AMGEN (to P.A.K.).

26 June 1998; accepted 4 September 1998

A Physical Map of 30,000 Human Genes

P. Deloukas,* G. D. Schuler, G. Gyapay, E. M. Beasley, C. Soderlund, P. Rodriguez-Tomé, L. Hui, T. C. Matise, K. B. McKusick, J. S. Beckmann, S. Bentolila, M.-T. Bihoreau, B. B. Birren, J. Browne, A. Butler, A. B. Castle, N. Chiannilkulchai, C. Clee, P. J. R. Day, A. Dehejia, T. Dibling, N. Drouot, S. Duprat, C. Fizames, S. Fox, S. Gelling, L. Green, P. Harrison, R. Hocking, E. Holloway, S. Hunt, S. Keil, P. Lijnzaad, C. Louis-Dit-Sully, J. Ma, A. Mendis, J. Miller, J. Morissette, D. Muselet, H. C. Nusbaum, A. Peck, S. Rozen, D. Simon, D. K. Slonim, R. Staples, L. D. Stein, E. A. Stewart, M. A. Suchard, T. Thangarajah, N. Vega-Czarny, C. Webber, X. Wu, J. C. Auffray, N. Nomura, J. M. Sikela, M. H. Polymeropoulos, M. R. James, E. S. Lander, T. J. Hudson, R. M. Myers, D. R. Cox, J. Weissenbach, M. S. Boguski, D. R. Bentley

A map of 30,181 human gene–based markers was assembled and integrated with the current genetic map by radiation hybrid mapping. The new gene map contains nearly twice as many genes as the previous release, includes most genes that encode proteins of known function, and is twofold to threefold more accurate than the previous version. A redesigned, more informative and functional World Wide Web site (www.ncbi.nlm.nih.gov/genemap) provides the mapping information and associated data and annotations. This resource constitutes an important infrastructure and tool for the study of complex genetic traits, the positional cloning of disease genes, the cross-referencing of mammalian genomes, and validated human transcribed sequences for large-scale studies of gene expression.

The ultimate gene map for an organism is the complete sequence of its genome, annotated with the beginning and ending coordinates of every gene. Construction of such sequence maps has become routine for simpler organisms with relatively small genome sizes (for example, 1 to 20 Mb), and public databases now contain 18 examples of such complete genomic sequences (*1*). For more complex organisms, such as mice and humans, with genome sizes in the 3-Gb range, complete and accurate genome sequences are still 5 to 10 years away (*2*, *3*). However, large quantities of preliminary data ("shotgun assemblies") are already available (*4*) and expected to grow rapidly (*5*). Both of these factors necessitate the construction of gene maps to support basic and applied research in mammalian biology and medicine, as well to aid in the analysis and interpretation of "unfinished" genome sequence data. Extensive libraries of expressed gene sequences (*6*, *7*), combined with physical mapping with radiation hybrid (RH) panels (*8*–*10*), have provided the information, infrastructure, and technology to produce such maps in an efficient and economical manner.

In 1994, an international consortium was formed to construct a human gene map in which cDNA-based sequence-tagged site (STS) markers were physically mapped and then integrated with the genetic map of polymorphic microsatellite markers (*11*). The initial report of this consortium in 1996 described a map of ~16,000 genes (*12*). A new map, reported here, represents a nearly 100% increase in gene density and map accuracy and may contain up to half of all human protein-coding genes. This map should be a valuable resource for the positional candidate cloning of complex (polygenic) disease loci, the construction of complete physical maps of chromosomes for genome sequencing, and comparative analysis of mammalian chromosome structure and evolution. Furthermore, sequence validation that occurs in the process of STS design and mapping creates a quality-assured gene sequence resource for "functional genomics" applications (*13*) such as the design and construction of large-scale gene expression arrays.

This new gene map consists of data from 41,664 STSs (Table 1). As in the previous map (*12*), they are based on 3′ untranslated regions of cDNAs. These STSs represent 30,181 unique genes. Markers were typed on the Genebridge4 (GB4) RH panel (39,886 cDNAs, 1641 microsatellite markers, and 13 telomeric markers), on the G3 RH panel (5013 cDNAs and 2091 microsatellites), or on both panels (1102 microsatellites). All GB4 data (Table 1) were, for the first time, merged into a single map and aligned with the G3 RH map and the genetic map (*11*) with the 1102 microsatellite markers that are common to all three maps. The integrated map is available at www.ncbi.nlm.nih.gov/genemap. In addition, two Web servers [one for each RH panel (*14*)] permit anyone to map a new marker relative to this map.