All-Atom Folding Simulations of the Villin Headpiece from Stochastically Selected Coarse-Grained Structures

Giacomo M. S. De Mori,[†] **Cristian Micheletti,***[‡] **and Giorgio Colombo***[†]

Istituto di Chimica del Riconoscimento Molecolare, CNR, Via Mario Bianco 9, 20131 Milano, and Scuola Internazionale Superiore di Studi Avanzati and INFM, Via Beirut 4, I-34014 Trieste, Italy

Received: May 24, 2004; In Final Form: June 28, 2004

We discuss the application of a novel efficient protocol for the numerical simulation of the folding dynamics of single domain proteins from the only knowledge of primary sequence. Our approach is based on the combination of a Monte Carlo (MC) coarse-grained evolution followed by all-atom molecular dynamics (MD) simulations in explicit solvent. The coarse-grained model simplifies the protein's energy landscape and allows it to evolve rapidly toward viable starting conformations for MD. A general fine-graining algorithm is then used to reconstruct the full atomic detail of the protein. All atom MD simulations in explicit water are then employed to investigate the protein's conformational evolution toward the native state. We discuss the application of this novel approach to the Villin headpiece, a widely studied test system for folding studies, for which we obtain and maintain an RMS deviation from the NMR structure of 2.4 Å for the core region and 3.7 Å for the whole protein. Finally, the analysis of the MC−MD trajectories provides valuable insight into important aspects of the folding process with regards to the appearance and docking of locally secondary structure elements.

The folding of proteins into their native structure lies at the heart of many important biological processes. It is believed that, for a large class of proteins, the knowledge of the mere primary sequence ought to allow the prediction of the protein's native structure.[1] Here we describe the results obtained by the application of a novel methodology for the theoretical study of the folding of single domain proteins from the only knowledge of sequence. Typical approaches to the problem rely either on the use of suitable sampling techniques to identify conformations of low free-energy or on the integration of the classical equations of motion for all of the atoms in the system. Our two-stage scheme is based first on a coarse-grained Monte Carlo (MC) evolution followed by all atom molecular dynamics (MD) simulations in explicit solvent. The effect of the coarse-grained part is to simplify the protein's energy landscape so to identify efficiently physically meaningful starting conformations for the MD. After bringing these structures into the realm of all-atom representations, explicit solvent MD introduces back the fine chemical details, which are ultimately responsible for driving the evolution toward the native state. The link between the two structural representations is a fine-graining algorithm which allows to reliably reconstruct the full atomic detail of the protein (see last paragraph and Supporting Information for details), using a library of previously generated protein fragments.

Coarse-grained models, where the protein is described as a chain of linked beads interacting via an effective potential have proved useful in the development of thermodynamic models of protein folding.[2] The dramatic reduction in the structural degrees of freedom allows to move through large regions of conformational space and hence evolve toward viable starting conformations for the MD evolution. All-atom MD simulations starting from these structures thus have the potential to explore parts of the phase space that would otherwise be inaccessible on the typical MD time-scale using random or totally extended starting conformations. The test system chosen for this study is the Villin headpiece HP36 (1VII.pdb),[3] a 36-residue protein with a well-defined tertiary structure used as a model in several theoretical investigations.[4]

The preliminary coarse-grained MC exploration of the free-energy landscape is achieved by describing the protein in terms of its Cα trace and of effective Cβ centroids. This is accompanied by a simplification of the energy functional which incorporates effective pairwise interactions among amino acids (KGS potentials),[5] knowledge-based constraints for backbone chirality, local propensities to form secondary motifs, and a term favoring their tertiary packing (see Supplementary Information). Within this simplified framework, the thermodynamics of HP36 was characterized by several MC evolutions at distinct temperatures. The MC dynamics entailed the use of pivot and crankshaft moves which preserve the length of the bonds, initially set to 3.8 Å, joining consecutive Cα centroids. As temperature is decreased, the protein undergoes a collapse, as signaled by the rapid decrease of both the radius of gyration, R_g, and the average system energy. The peak in the specific heat in correspondence of the collapse temperature, T_c, is further associated with significant fluctuations in energy reflecting the coexistence of rather swollen and globular conformations (Figure 1). The latter ones typically possess local secondary elements and are further compactified at lower temperatures. The protein at T_c is thus poised to collapse into compact conformations with nontrivial secondary content. Thus, the structures encountered in the MC trajectory at T_c represent attractive candidates for all-atom MD evolution for several reasons: (1) secondary elements are typically formed, (2) the structures are not unnaturally compact, and (3) the conformational variability is

* To whom correspondence should be addressed. (C.M.) E-mail: michelet@sissa.it. Telephone: ++39-040-2240456. (G.C.) E-mail: giorgio.colombo@icrm.cnr.it. Telephone: ++39-02-28500031.

[†] Istituto di Chimica del Riconoscimento Molecolare.

[‡] Scuola Internazionale Superiore di Studi Avanzati and INFM.

10.1021/jp0477699 CCC: $27.50
Published on Web 07/23/2004

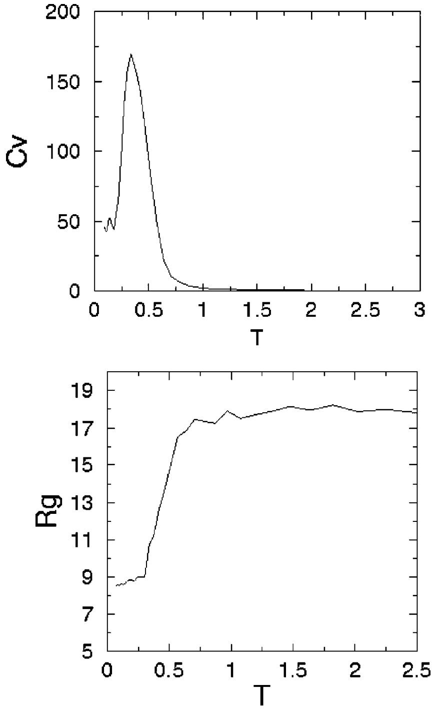

Figure 1. Specific heat (C_v, top) and radius of gyration (R_g, bottom) as a function of the temperature of the coarse-grained MC model.

such that significantly different structures can be picked (the average RMSD between any pair of structures sampled at T_c being 5.9 ± 1.2 Å).

Seven different uncorrelated coarse-grained conformations were thus chosen at T_c and used, after the fine-graining, for the subsequent all-atom MD (see Methods and Supplementary Information).

Under the assumption that the free-energy landscape retains the relevant features of the true one, the all-atom MD simulations starting from the sampled conformations are expected to have significant advantages over, e.g., those starting from fully extended protein configurations which can be affected by significant lag-phases.[6] The scheme adopted here therefore aims at extending the reach of ordinary MD simulations by using simple physicochemical criteria to identify the marginally compact starting conformations in which partial formation of secondary and tertiary interactions has taken place. Thus the approach has a different spirit from that followed in other two-stage approaches (notably the pioneering ones of Vieth et al.[7] and Liwo et al.[8]) where the fine-graining step was aimed at perturbing or refining the coarse-grained structures which minimized a given energy functional.

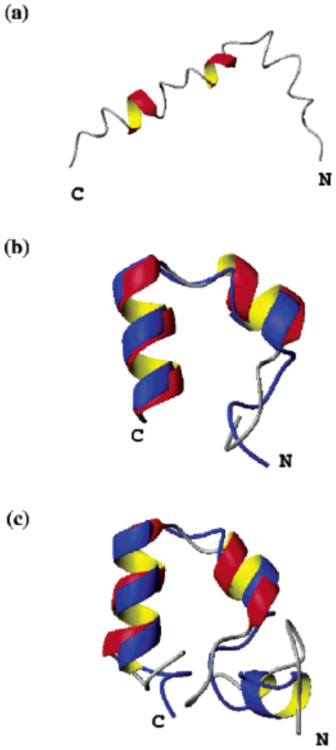

Figure 2. (a) Starting structure of simulation F1. Superposition of the representative structure for F1 (red) with 1VII (blue) over the (b) core region. (c) and the whole protein.

TABLE 1: Summary of the Average Structural Features Observed in the Last 20 ns of Simulations F1 to F7 and in the Last 10 ns for Simulations Ext1 to Ext 3

simulation	RMSD core 9−32 (Å)	RMSD all (Å)	R_{gyr} (Å)
F1	3.0 ± 0.2	4.1 ± 0.2	9.2 ± 0.1
F2	6.1 ± 0.3	8.5 ± 0.2	9.3 ± 0.1
F3	6.9 ± 0.2	7.0 ± 0. 3	1.00 ± 0.2
F4	5.4 ± 0.3	8.2 ± 0.5	9.2 ± 0.4
F5	5.1 ± 0.2	6.0 ± 0.3	9.5 ± 0.3
F6	7.2 ± 0.3	8.1 ± 0.4	9.8 ± 0.3
F7	5.4 ± 0.2	7.4 ± 0.2	9.6 ± 0.1
Ext 1	8.1 ± 2.4	9.2 ± 3	11.8 ± 3
Ext 2	8.0 ± 1.7	9.5 ± 2.8	11.4 ± 3.4
Ext 3	8.3 ± 1.9	9.4 ± 2.1	10.4 ± 3.9

The all-atom conformations obtained from the fine-graining procedure were then solvated with SPC water,[9] energy minimized, equilibrated and evolved at 300K for 50ns using the Gromos96[9] force field, PME treatment of electrostatics,[9] and the GROMACS program.[9] To check the viability of the force field at this temperature, and to produce a term of reference, a 50 ns trajectory was produced starting from the minimized average NMR structure taken from PDB (1VII). The evolution shows a marked stability of R_g and of the whole protein structure. The average root mean square deviation (RMSD) value over the core (residues 9−32 as defined in ref 4) is around 2 Å, consistently with what observed by Duan and Kollman[4] The three native helical elements, H1 (residues 4−8), H2 (15−18), and H3 (23−30), are well maintained throughout the simulation. For the sake of brevity, of the seven simulations carried out starting from the coarse-grained structures collected at T_c (F1 to F7), we will concentrate on three of them representative of the overall observed dynamic behavior, namely F1, F4, and F7. In Table 1, we summarize the results concerning the radius of gyration and RMSD deviations from the native reference conformation (the minimized average NMR structure) recorder in the last stages of the various trajectories.

Trajectory F1 starts from a rather open configuration but with some helical content in regions H2 and H3 (see Figure 2a). After a short equilibration time, it undergoes a rapid compaction as shown by the fast decrease in R_g values. Not only is this collapse rapid (10 ns), but the initial helical segments grow to the full native extension of H2 and H3 while also achieving their correct tertiary packing. The core RMSD with respect to the minimized average NMR structure often attains values as low as 2.8 Å and stabilizes around a mean value of 3.0 Å (Figures 2b and 3a).

The minimum RMSD calculated over all NMR models of the protein was 2.4 Å and 3.7 Å for the core region and the whole protein, respectively. The superposition of the average NMR-minimized structure with the representative structure obtained from the most populated structural cluster obtained from cluster analysis of trajectory F1 is displayed in Figure 2b,c.

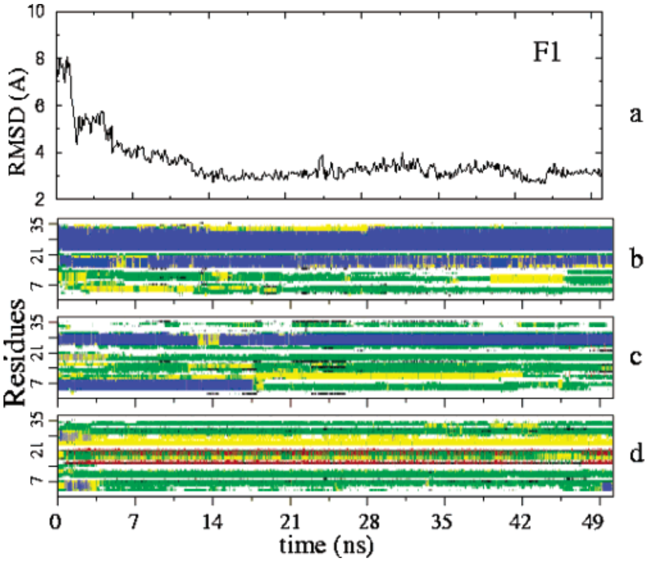

Figure 3. Time evolution of (a) core RMSD with respect to the minimized average NMR structure in simulation F1and of the secondary structure content in F1 (b), F4 (c), and F7 (d). Color code: blue, helix; yellow, turn; green, bend; red, β-strand; white, disordered structure.

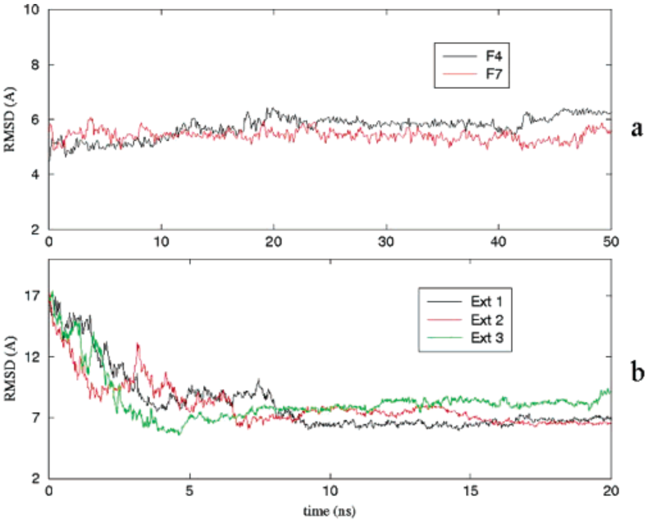

Figure 4. Time evolution of the core RMSD with respect to the minimized average NMR structure in simulations started (a) from two other structures picked at T_c and (b) from fully extended conformations.

This result constitutes a significant advancement over the pioneering study of Duan and Kollman where the representative structure of their 1-μs-long simulation was found to be at 4 and 5.7 Å RMSD from the NMR reference for the core and the whole protein, respectively.[4] The advantage of the present approach, which overall involved a fraction of the time-span simulated in ref 4, is that valuable insight about the folding dynamics can be obtained by the analysis and comparison of the evolution of the various starting structures. Trajectory F4, for example, starts from a conformation with good helical content in regions H1 and H3 (Figure 3c). In about 10 ns, it attains a satisfactory native similarity for the core (4.3 Å RMSD, Figure 4a) which is largely lost in the subsequent evolution. Analysis of the secondary content allows to associate this degradation with the disappearance of the helix H1 in favor of the formation of two contacting strands involving residues 2−7 and 15−20. This conformation remains stable for the subsequent dynamical evolution suggesting its possible importance as a local free-energy minimum. Interestingly, the tendency to form β-sheets is also observed in other trajectories, as F7, where no native progress is recorded (the average RMSD being 6 Å, Figure 4a) and that persistently display contacting extended segments organized in an overall compact structure.

Besides these trajectories, we have carried out runs starting from fully extended conformations with the purpose of quantifying the degree of nativeness of the structures reached from completely open conformations (Figure 4b). In all of the cases examined, the extended structures rapidly collapse to compact states in which, unlike cases F1−F7, no secondary or tertiary structure ordering is apparent. These runs were stopped after 20 ns since MD at 300 K was unable to induce, on this time scale, the conformational changes required to break the compact and disordered globular structures obtained. Furthermore, the larger number of water molecules required in the extended runs made the dynamical evolution so slow that the three 20 ns long simulations required approximately the same CPU time of the seven 50 ns long trajectories F1−F7.

This fact testifies the usefulness and efficiency of the hybrid approach used here. In fact, the preliminary MC exploration brings the considerable advantage of starting the all-atom evolution from structures that have both a nontrivial secondary content and a good compactness.[6] Although the coarse-grained force field may be imperfectly parametrized, a fraction of the sampled structures have a good native secondary content. When the local secondary elements are sufficiently close to one another, as in F1, the all-atom evolution can rapidly progress toward the native state.

The analysis of the evolution of other structures shows the propensity of certain sites to attain extended configurations which impair the progress toward the native basin. This is clearly manifested by run F4 where the loss of the promising initial native similarity (the core RMSD being 4.5 Å) is paralleled by the formation of a turn involving residues 8−10 and the pairing of beta-like structures involving residues 2−7 and 10−15. The same residues displayed the tendency to form beta structures in two other trajectories. For the case of simulation F3, a different mechanism seems to be responsible for the inability to evolve rapidly toward the native basing despite the correct formation of the three native helical segments. In this case, in fact, the helices assemble in a non-native manner due to the formation of a hydrophobic core (involving Phe11, Leu21, Trp24, and Leu29) which remains stable during the simulated time-span. The results of simulation F1, on the other hand, show that the rapid attainment of the folded conformation is correlated with the presence and correct favorable orientation of helices H2 and H3. The latter can thus rapidly dock onto each other determining the formation of favorable hydrophobic contacts of Phe18, with the alkylic side chains of Gln26, Leu29, and Lys30 which drive the system into the native basin. The key role played by the Phe residues in the MD trajectories is consistent with the recent experimental findings of ref 10 which observed a loss of ordered secondary structures in HP36 Phe to Leu mutants.

Although the present MD trajectories are not aimed at providing a thorough characterization of the protein free-energy landscape (e.g., in terms of intermediates or transition state(s)), they strongly support the fact that the fast progress toward the folded state appears to require the harmonious interplay of local and nonlocal interactions. This picture is consistent with several theoretical and experimental observations which identify the crucial step of the folding process with the establishment of a minimal set of short- and midrange native contacts.[1,4,11]

To summarize, we have followed the all-atom dynamical evolution of the Villin headpiece starting from several configurations obtained by a coarse exploration of the system energy landscape. The seven simulated trajectories included a notable instance where the folded configuration of the protein core was reached and maintained (see Figures 2b,c and 3a,b). Furthermore, the significant secondary content and organization found in all starting structures resulted in interesting dynamical evolutions that, even when not progressing toward the folded state, convey valuable information on the folding process, as the trapping mechanism associated with the formation of contacting strands. The protocol used here has been kept as general and unbiased as possible and may be further extended, e.g., at the level of the coarse-grained model and/or of the selection criteria of the starting structures. Therefore, the proposed strategy ought to be applicable to other instances of short proteins with significant advantages over the use of extended structures as the default unbiased starting point of all-atom dynamical simulations with explicit solvent.

Reconstruction Procedure. The algorithm that reconstructs the full atomic detail of the protein starting from its preassigned Cα trace is based on the use of a library of protein fragments built from 100 representative structures resolved by NMR. These structures were parsed into fragments of four consecutive residues retaining the whole atomic detail of the mainchain and of the side chains of the middle residues. For each set of four consecutive Cα's in the Cα trace, one finds the best superimposable fragment in the library and assigns the middle backbone to the protein to be reconstructed (the first and last peptidic planes are treated separately). Finally, the side chain of a given residue, R, is obtained by first considering only the set of fragments where one of the middle residues, R', is of the same type as R. Next, by putting in correspondence R and R', the side chain of R is assigned (again after an optimal roto-translation) from the fragment providing the best superposition with the reconstructed backbone. As shown in the Supplementary Information, this procedure allows us to place 70% of the heavy atoms within 1 Å of the crystallographic positions (and 62% within only 0.5 Å).

Supporting Information Available: Details of materials and methods. Coarse-grained model and Monte Carlo evolution and the reconstruction procedure. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) (a) Anfinsen, C. B. *Science* **1973**, *181*, 223. (b) Anfinsen, C. B.; Scheraga, H. A. *Adv. Protein Chem.* **1975**, *29*, 205−300.

(2) (a) Levitt, M. A. *J. Mol. Biol.* **1976**, *104*, 59−107. (b) Sali, A.; Shakhnovich, E.; Karplus, M. *Nature* **1994**, *369*, 248−251. (c) Lee, J.; Liwo, A.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 2025−2030. (d) Harrison, P. M.; Chan, H. S.; Prusiner, S. B.; Cohen, F. E. *J. Mol. Biol*. **1999**, *286*, 593−606. (e) Klimov, D. K.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 2544−2549.

(3) McKnight, C. J.; Matsudaira, P. T.; Kim, P. S. *Nat. Struct. Biol*. **1997**, *4*, 180−184.

(4) (a) Duan, Y.; Kollman, P. A. *Science* **1998**, *282*, 740−744. (b) Pande, V. S.; Baker, I.; Chapman, J.; Elmer, S. P.; Khaliq, S.; Larson, S. M.; Rhee, Y. M.; Shirts, M. R.; Snow, C. D.; Sorin, E. J.; Zagrovic, B. *Biopolymers* **2003**, *68*, 91−109. (c) Shen, M.; Freed, K. F. *Proteins* **2002**, *49*, 439−445. (d) Jang, S.; Kim, E.; Shin, S.; Pak, Y. *J. Am. Chem. Soc*. **2003**, *125* (48), 14841−14846.

(5) Kolinski, A.; Godzik, A.; Skolnick, J. *J. Chem. Phys*. **1993**, *98*, 7420−7433.

(6) Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 14122−14125.

(7) Vieth, M.; Kolinski, A.; Brooks, C. L.; Skolnick, J. *J. Mol. Biol.* **1994**, *237*, 361−367.

(8) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *Prot. Sci.* **1993**, *2*, 1715−1731.

(9) (a) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. In *Intermolecular Forces*; Pullman, B., Ed.; D. Reidel Publishing Company: Dordrecht, The Netherlands, 1981; pp 331−342. (b) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular Simulation: The GROMOS96 Manual and User Guide*; Hochschuleverlag AG an der ETH Zürich: Zürich, Switzerland, 1996. (c) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089−10092. (d) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Mod*. **2001**, *7* (8), 306−317.

(10) Tang, Y.; Rigotti, D. J.; Fairman, R.; Raleigh, D. P. *Biochemistry* **2004**, *43*, 3264−3272.

(11) (a) Shortle, D.; Ackerman, M. S. *Science* **2001**, *293*,487−489. (b) Debe, D. A.; Carlson, M. J.; Goddard, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 2596−2601.