JCTC Journal of Chemical Theory and Computation

Protein Folding Pathways Revealed by Essential Dynamics Sampling

Daniele Narzi,[†,§] Isabella Daidone,*[,†,||] Andrea Amadei,[‡] and Alfredo Di Nola[†]

Department of Chemistry, University of Rome 'La Sapienza', P.le Aldo Moro 5, 00185 Rome, Italy, and Dipartimento di Scienze e Tecnologie Chimiche, University of Rome 'Tor Vergata', via della Ricerca Scientifica 1, I-00133 Rome, Italy

Received May 8, 2008

Abstract: The characterization of the protein folding process represents one of the major challenges in molecular biology. Here, a method to simulate the folding process of a protein to its native state is reported, the essential dynamics sampling (EDS) method, and is successfully applied to detecting the correct folding pathways of two small proteins, the all-β SH3 domain of Src tyrosine kinase transforming protein (SH3) and the α/β B1 domain of streptococcal protein G (GB1). The main idea of the method is that a subset of the natural modes of fluctuation in the native state is key in directing the folding process. A biased molecular dynamics simulation is performed, in which the restrained degrees of freedom are chosen among those obtained by a principal component, or essential dynamics, analysis of the positional fluctuations of the Cα atoms in the native state. Successful folding is obtained if the restraints are applied only to the eigenvectors with lowest eigenvalues, representing the most rigid quasi-constraint motions. If the essential eigenvectors, the ones accounting for most of the variance, are used, folding is not successful. These results clearly show that the eigenvectors with lowest eigenvalues contain the main mechanical information necessary to drive the folding process, while the essential eigenvectors represent the large concerted motions which can occur without folding/unfolding the protein.

1. Introduction

Understanding protein folding mechanisms represents one of the major aims of biophysics and molecular biology. Information on the sequence of conformational steps that lead to the native structure from denaturated polypeptide chains is fundamental to shed light on protein folding mechanisms, on the effects of different physicochemical conditions and mutations. Many experimental and theoretical approaches for the study of protein folding have been developed.[1−8] Computational methods represent a valid tool in order to obtain atomic details of such a process, and molecular dynamics (MD) simulations are among the most used ones.[9−12] A limitation encountered using MD simulations is due to the time scale accessible to this methodology that is not comparable with the time scale of most folding processes (ms-s). At present, with standard MD simulations this process can be well simulated only for short peptides[9,11,13,14] but is still beyond reach for globular proteins. To overcome this problem different MD techniques have been developed.

The simplest approach is to perform high-temperature MD simulations starting from the native structure to study the unfolding process.[15−22] In some instances, by considering the unfolding as the reverse of folding, information on the folding process is inferred from the high-temperature unfolding simulations.[20−22] A more sophisticated approach makes

* Corresponding author e-mail: Isabella.Daidone@iwr.uni-heidelberg.de.

† University of Rome 'La Sapienza'.

§ Current address: Theoretical & Computational Membrane Biology, Center for Bioinformatics Saar, Universität des Saarlandes, D-66041 Saarbrücken, Germany.

|| Current address: Interdisciplinary Center for Scientific Computing, University of Heidelberg, Im Neuenheimer Feld 368, 69120 Heidelberg, Germany.

‡ University of Rome 'Tor Vergata'.

10.1021/ct800157v CCC: $40.75
Published on Web 10/21/2008

use of unfolding simulations followed by the calculation of the free energy of the folding process at $T = 300$ K, by means of the umbrella sampling method, along the previously determined path. Unfolding is performed by high temperature[23,24] or by applying a harmonic potential to different reaction coordinates, such as the radius of gyration[25] and the end-to-end distance.[26,27] The previously reported approaches are based on the hypothesis that the folding process at 300 K follows the same path of the unfolding process performed with high temperature or with a harmonic potential, but the issue of whether unfolding simulations are representative for the folding process is still open.[28,29]

Different from the previous methods, in 'targeted molecular dynamics' (TMD) simulation,[30] a folding simulation is performed along a path not previously determined. This is accomplished by applying a harmonic, time-dependent, restraint on each atom to continuously decrease the all-atom root-mean-square deviation from the native state. Other methods make use of simplified molecular models in order to gain computation time by neglecting details. This category includes the widely used so-called lattice models.[31,32] Additionally, accurate prediction of native three-dimensional protein structures could be reached using semiempirical database-driven prediction methods.[33−35]

Here, the essential dynamics sampling (EDS) method[36,37] is used to simulate protein folding. Starting from an unfolded structure, a usual MD simulation step is performed. The new structure is accepted only if its distance from the native structure does not increase, otherwise, it is projected onto the closest configuration having the same distance to the native conformation as the structure before the MD step. The distance is calculated in a configurational subspace defined by a set of generalized coordinates obtained by a principal component, or essential dynamics, analysis[38−40] of a native-state equilibrium simulation. Hence, correct folding can be obtained by using only a small fraction of the degrees of freedom of the protein to bias the MD simulation toward its native conformation.

Due to the absence of any restraining potential, as used for example in TMD, the EDS method does not force the system to overcome barriers higher than a few kT. Thus, the protein is not allowed to undergo major unfolding if incorrect packing leads to kinetic traps or, in other words, to off-pathway intermediates. In this sense the method is somewhat similar to the CONTRA MD[41] but differs mainly in the choice of the reaction coordinates which, in the present case, are chosen so to contain information on the dynamical properties of the native state and might, hence, represent better candidates than reaction coordinates often used such as the radius of gyration or the root-mean-square deviation from the native structure.

The EDS method was successfully applied to the folding process of cytochrome *c*,[42] an all-α protein. In the present work, the method is further extended, and its ability in reproducing the native conformation and the known folding steps in proteins with different topologies, namely the all-β and α/β motives, is verified. The model systems used are the SH3 domain of Src tyrosine kinase transforming protein (SH3) and the B1 domain of streptococcal protein G (GB1). The src-SH3, a 56-residue all-β protein, was largely investigated by MD simulations,[43−46] and, in agreement with experimental data,[47,48] the folding transition state is found to be characterized by the presence of the central three-stranded β-sheet, whereas the formation of the hydrophobic sheet, consisting of the two terminal strands, is observed in the last stage of the folding process. The GB1, a 56 residues α/β protein, has been shown to populate an intermediate state along its folding process with native-like structural elements involving one of the four strands, namely the β3 strand.[49,50]

The results of the EDS folding simulations performed here show that in SH3 the central three-stranded β-sheet precedes the whole structure formation and in GB1 the native-contacts formation of the β3 strand is a prerequisite for a correct folding. These results are in agreement with experiments,[47−50] thus assessing the predictive capabilities of the method.

2. Methods

2.1. Molecular Dynamics Simulations. All MD simulations were performed using the GROMACS software package and the Gromos87 force field[51] with modification as suggested by van Buuren et al.[52] In both cases the proteins were solvated with water in a periodic cubic box of dimensions $57.0 \times 57.0 \times 57.0$ Å. The simple point charge[53] water model was used. Neutralization of the total charge of the system was obtained by replacing 3 and 4 molecules of water with 3 and 4 Na ions for SH3 and GB1, respectively. The SHAKE algorithm[54] was used to constrain all bond lengths. A time step of 2 fs was used for numerical integration. The isokinetic temperature coupling[55] was used to keep the temperature constant. The long-range electrostatic interactions were treated with the particle mesh Ewald method[56] using a $48 \times 48 \times 48$ grid combined with a fourth-order B-spline interpolation to compute the potential and forces in between grid points, whereas the short-range electrostatic interactions were treated with a nonbonded pair-list cutoff of 9.0 Å.

2.2. Native-State Simulations and Essential Dynamics (ED) Analysis. For both SH3 and GB1 a 5000 ps long MD simulation of the native state was performed at room temperature ($T = 300$ K) in the NVT ensemble (at a liquid density of 55.32 mol/L). The starting structures were taken from the NMR structure (PDB entry 1srl)[57] for SH3 and from the 2.07 Å resolution refined crystal structure (PDB entry 1pga)[58] for GB1. From the equilibrated portion of the native-state trajectory (beyond 200 ps) the covariance matrix of the positional fluctuations of the Cα carbon atoms was built up and diagonalized. The procedure yields new axes (eigenvectors), representing the directions of the concerted motions. The corresponding eigenvalues give the mean square positional fluctuation for each direction.[38,39] 168 eigenvectors were obtained for each protein, corresponding to the number of degrees of freedom of the Cα carbon atoms. Sorting the eigenvectors by the size of the eigenvalues shows that the configurational space can be divided in a low dimensional (essential) subspace (the first 10−15 eigenvectors in the present proteins) in which most of the positional fluctuations are confined ($\approx$ 60−70% of the total variance) and a high

dimensional (near-constraint) subspace in which small-amplitude fluctuations occur.

2.3. Essential Dynamics Sampling (EDS). The essential dynamics sampling technique[36,37,59] can be used to decrease the distance of a given structure from a reference structure in a space defined by a subset of eigenvectors as obtained by the ED analysis of the native-state MD simulation (see the Results section for the choice of the set of eigenvectors used in the present work).

In the EDS simulation a usual MD simulation step is performed starting from an unfolded conformation; at each step the distance from the reference conformation (the crystal or NMR structure in the present cases) is calculated in the chosen subspace. If this distance does not increase, the new conformation is accepted. Otherwise, the coordinates (in the chosen subspace) are radially corrected in order to keep the position onto the hypersphere centered on the reference conformation, with a radius given by the distance from the reference in the previous step. This correction step is performed using a nonstationary holonomic constraint in the chosen subspace ξ

$$G(\xi(t+\Delta t);t+\Delta t) = |\xi(t+\Delta t) + \Delta\xi_c - \xi_0|^2 - |\xi(t) - \xi_0|^2 = 0 \quad (1)$$

where $\xi(t)$ and $\xi(t+\Delta t)$ are the unconstrained positions at time t and $(t+\Delta t)$, respectively, $\Delta\xi_c$ is the correction for the application of the constraint, and ξ_0 is the reference position (the crystal or NMR structure). Eq 1 does not suffice to solve for $\Delta\xi_c$ in a unique way. To obtain a unique solution, we add the requirement that $|\Delta\xi_c|^2$ is minimized. This is achieved using one Lagrangian multiplier:

$$\Delta\xi_c^i - \lambda\frac{\partial G}{\partial\Delta\xi_c^i} = \Delta\xi_c^i - 2\lambda(\xi^i(t+\Delta t) + \Delta\xi_c^i - \xi_0^i) = 0 \quad (2)$$

and therefore

$$\Delta\xi_c^i = \frac{2\lambda}{1-2\lambda}(\xi^i(t+\Delta t) - \xi_0^i) \quad (3)$$

Using eq 3 and eq 1 λ can be expressed as a function of $\xi(t)$, $\xi(t+\Delta t)$, and ξ_0. This value of λ, and the corresponding $\Delta\xi_c$, is then used to correct $\xi(t+\Delta t)$ to fulfill the constraint with the least perturbation.

2.4. Unfolding/Folding Simulations. Starting from two different structures extracted from the native-state MD simulations at t = 2000 ps and t = 3000 ps, two high temperature unfolding simulations of 3500−4000 ps were performed for each protein. The temperature was kept at a value of 500 K, and the system was coupled to a pressure bath at a value of 1 bar. It has to be pointed out that these conditions are not meant to represent a real unfolding process, and the corresponding trajectories are not used for analysis purpose. They are rather meant as a computational procedure to generate a large number of denatured structures to be used as starting points in the folding simulations. For the GB1 protein a further simulation of 5000 ps was performed coupling the residues corresponding to the β3 strand (GLY41-ASP47) at a temperature of 300 K and the remaining residues at 500 K (see the Results section for the justification of this simulation).

Six and twelve protein structures for the SH3 and the GB1, respectively, were extracted from these unfolding simulations and used as starting structures in the EDS folding simulations. The selected structures are characterized by high values of root-mean-square deviation with respect to either the NMR or the crystal structure and high radius of gyration. Nevertheless, they retain some degree of secondary structure (see the Results section). Experimental and computational methods have demonstrated that even under strong denaturing conditions unfolded structures retain a residual native-like secondary structure.[60,61] Therefore, we believe that the unfolded structures used in the present work, which in fact contain information from the native starting conformation, are good candidates as representative structures of the unfolded ensemble.

The starting unfolded structures were solvated in water and equilibrated for 10 ps at a temperature of 300 K and a pressure of 1 bar. The folding simulations were then performed in the NVT ensemble at room temperature (T = 300 K). A slightly different procedure is used in the present work, with respect to the one previously reported:[42] to allow a local increase of the distance from the reference, each 10 ps of EDS simulation is followed by 10 ps of unbiased MD simulation.

3. Results

3.1. EDS Procedure. Preliminary analyses were performed to assess the relevance of using different sets of the native eigenvectors, accounting for the Cα carbon atoms fluctuations, in the biasing procedure of the folding simulations (only results for the GB1 are reported here since similar results are also obtained for SH3).

Starting from an unfolded conformation (structure RUN1 in Figure 1) three initial folding simulations were performed using all the Cα eigenvectors, the high-variance essential eigenvectors (the first 13), and the low-variance eigenvectors (the last 155) - $RUN1_{all}$, RUN1′, and RUN1, respectively, in Table 1. At the end of both simulations that included the essential eigenvectors a compact structure is reached, but almost no secondary and tertiary structure is recovered (see $RUN1_{all}$ and RUN1′ in Table 1).

In order to characterize the two sets of the Cα eigenvectors, i.e., providing and not providing correct folding, the nature of the associated motions was investigated. For this purpose, the overall displacement of the Cα atoms belonging to a given secondary-structure element was decomposed into internal motions, i.e., occurring within the secondary structure, and roto-translational motion, i.e., of the secondary-structure element with respect to its Cα centroid. An example for the β3-β4 sheet of GB1 is reported in Figure 2. The results make evidence that the last, i.e., with the lowest eigenvalues, 150−155 eigenvectors (out of 168) mostly represent internal collective vibrations, i.e., within the β-sheet, whereas the essential eigenvectors (the first 10−20) mainly provide roto-translational motions of the β-sheet.

These results show that the quasi-constraint, low-variance eigenvectors, that were shown here to represent in the folded protein the smallest vibrations within each secondary struc-

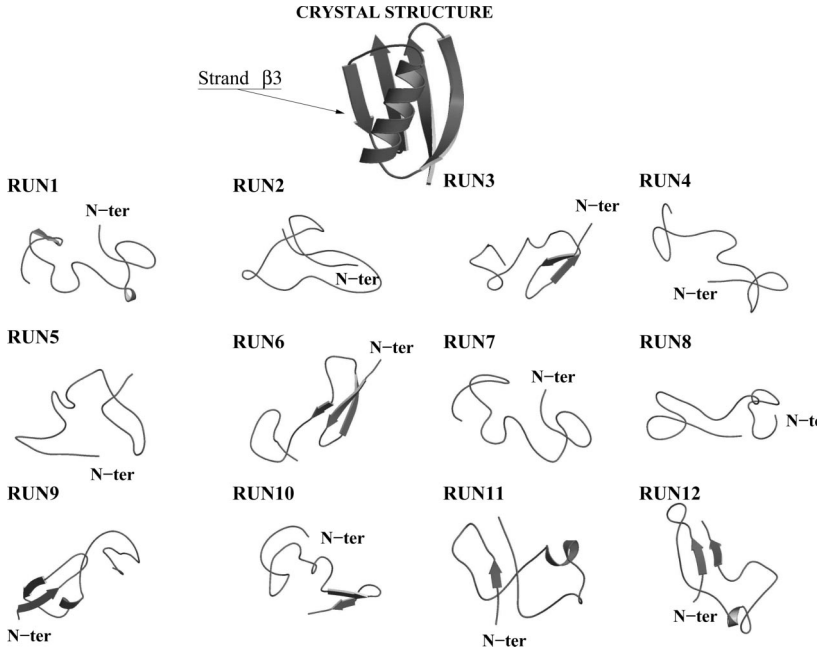

Figure 1. GB1. Backbone conformation of the crystal structure and of the twelve structures used as starting structures in the EDS folding simulations.

Table 1. GB1: Starting and Final Radius of Gyration (Rg_i, Rg_f), Backbone Root-Mean-Square Deviation ($RMSD_i$, $RMSD_f$) and Native Contact Fraction (ρ_i, ρ_f) with Respect to the Crystal Structure, Number of Residues in β-Structure (N_{β_i}, N_{β_f}) and in α-Structure (N_{α_i}, N_{α_f}) in the Folding Simulations[a]

SIM	Rg_i (nm)	Rg_f (nm)	$RMSD_i$ (nm)	$RMSD_f$ (nm)	ρ_i	ρ_f	N_{β_i}	N_{β_f}	N_{α_i}	N_{β_f}
crystal	1.05		–		–		24		14	
NatGB1	1.05(0.01)		0.12(0.02)		0.91(0.02)		23(1)		14(1)	
RUN1$_{all}$	1.28	1.07	1.14	0.15	0.25	0.51	0	10	0	7
RUN1′	1.28	1.06	1.14	0.15	0.25	0.48	0	8	0	5
RUN1	1.28	1.07	1.14	0.13	0.25	0.89	0	21	0	15
RUN2	1.31	1.05	1.19	0.10	0.18	0.92	8	22	4	15
RUN3	1.27	1.05	0.97	0.14	0.40	0.84	6	23	9	14
RUN4	1.47	1.05	1.21	0.12	0.28	0.88	3	16	2	15
RUN5	1.30	1.09	1.15	0.32	0.25	0.61	0	7	5	0
RUN6	1.37	1.03	1.03	0.15	0.32	0.86	15	16	0	14
RUN7	1.27	1.04	1.12	0.13	0.27	0.79	2	19	6	11
RUN8	1.32	1.04	1.16	0.20	0.18	0.75	0	14	2	14
RUN9	1.31	1.02	1.02	0.17	0.37	0.85	6	12	4	15
RUN10	1.19	1.06	1.24	0.27	0.27	0.71	15	4	0	15
RUN11	1.24	1.05	1.10	0.15	0.44	0.85	7	20	0	13
RUN12	1.28	1.05	1.13	0.08	0.35	0.90	10	24	0	15
RUN4′	1.47	1.07	1.21	0.14	0.28	0.82	3	21	2	15
RUN8′	1.32	1.06	1.16	0.13	0.18	0.81	0	20	2	14

[a] The final values in the folding simulations are averaged over the last 100 ps of each simulation. The values for the native-state trajectory (NatGB1) are averaged on the equilibrated part (200−5000 ps) with standard deviations in parentheses. The number of Cα eigenvectors used in the EDS procedure is as follows: all in RUN1$_{all}$; the first 13 in RUN1′; the last 155, i.e., the last 90%, in RUN1-RUN12; the last 90% of the eigenvectors calculated including not only all the Cα atoms but also the side-chain atoms of residues 41−47 in RUN4′ and RUN8′.

ture element, contain the proper mechanical information for the folding process, whereas the essential eigenvectors represent the large collective motions which can occur without folding/unfolding the protein.

It should be noted that in the previous study on the cytochrome *c*,[42] an all-α protein, a correct folding of the protein was obtained performing EDS folding simulations on a smaller space with respect to the GB1 and SH3, i.e., the last 30% of the eigenvectors versus the last 90% used here. When only the last 30% of the eigenvecors was used for GB1 and SH3, folding was not successful (data not shown). We assign this difference to the fact that β or α/β folds, such as SH3 and GB1, are characterized by higher contact order with respect to α topologies, such as cytochrome *c*, and hence the main mechanical information necessary for folding is distributed over a larger number of degrees of freedom.

In what follows we will perform different independent folding simulations using the last 155 eigenvectors for the GB1 and the last 160 for SH3.

3.2. GB1. The main structural properties of the native-state MD simulation at 300 K (NatGB1) are reported in Table 1. The data show a good agreement with the crystal structure. The values of the radius of gyration (Rg), root-mean-square

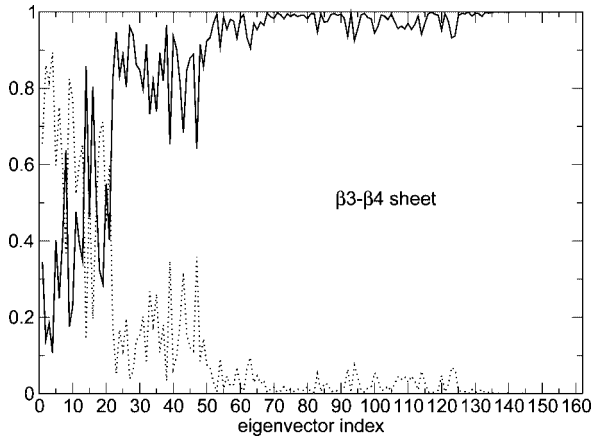

Figure 2. Fraction of internal (solid line) and roto-translational, with respect to the Cα atoms centroid (dotted line), displacements of the Cα atoms due to the motion along each eigenvector for the β3-β4 sheet of GB1. The following procedure was used: configurations of the secondary structure element of interest, as obtained by the filtered motion of a given eigenvector, were least-squares fitted to the corresponding average configuration. The mean square fluctuation recalculated for the given secondary-structure element after this procedure provides the internal-motion contribution to the total mean square fluctuation of the secondary structure element due to the eigenvector motions, while the residual fluctuation is ascribed to the roto-translational motion of the secondary structure element.

deviation (RMSD), % of native contacts (ρ) with respect to the crystal structure, and number of residues in β- and α-structure (N_β and N_α) of the ten unfolded structures used as starting points for the folding simulations are reported in Table 1 as well. All the starting structures are characterized by high RMSD and Rg and low native contacts and secondary structure contents. The corresponding conformations are shown in Figure 1, together with the crystal structure.

To simulate the folding process, the EDS was performed for a time range of 3000−5000 ps for each starting structure in a subspace defined by the last 155 eigenvectors of the covariance matrix of the Cα positional fluctuations (see the 'EDS Procedure' section). The final structural properties, averaged over the last 100 ps of each folding simulation, are reported in Table 1. Although from the table it is not completely clear which simulations are really successful, further analyses (vide infra) suggest that three simulations (RUN1−3) out of ten were successful.

The side-chain RMSD, with respect to the crystal structure, averaged on the three EDS simulations providing the correctly folded structures and on the last 100 ps of each simulation, is reported in Figure 3. A good agreement with the RMSD calculated on the native-state trajectory, reported in the same figure, can be observed. This result shows that, although the constraint applied in EDS accounts only for Cα atoms, the correct conformation of the side chains was obtained in EDS folding simulations.

The analysis of the trajectories shows that to achieve the final correct folding of the β3-β4 sheet, the TRP43 and TYR45 side chains (belonging to the β3 strand shown in

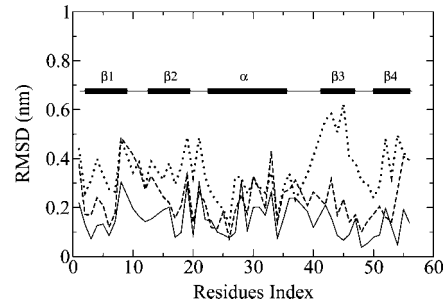

Figure 3. GB1. Side-chain RMSD with respect to the crystal structure. Solid line: average on the equilibrated part of the native-state simulation (200−5000 ps). Dashed line: average over the last 100 ps of the three correctly folded simulations (RUN1−3). Dotted line: average over the last 100 ps of the seven not correctly folded simulations (RUN4−10).

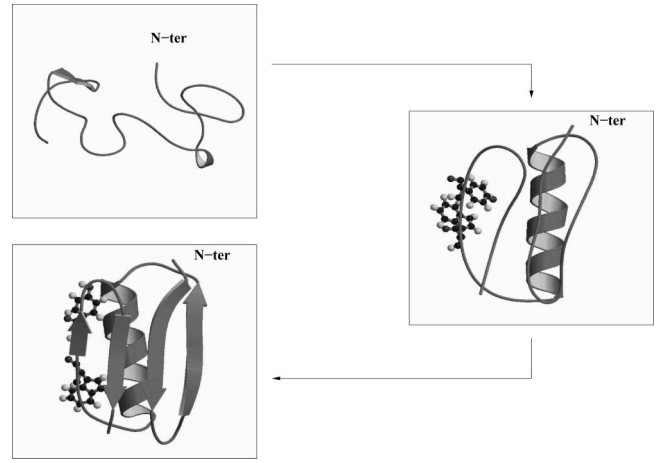

Figure 4. GB1. Backbone conformations at $t = 0$ ps, 984 ps, 4000 ps for RUN1. The side-chain orientations of TRP43 and TYR45 are also reported at $t = 984$ ps and $t = 4000$ ps.

Figure 1) need to be oriented toward the α-helix. As an example in Figure 4 we report representative structures along RUN1, the initial and final conformations, and a conformation observed at an intermediate time ($t = 984$ ps), in which the TRP43 and TYR45 side chains, highlighted in the figure, both point at the interface with the α-helix.

Conversely, at the end of the seven nonsuccessful simulations the region corresponding to the β3 strand (residues 41−47) is characterized by very high values of the side-chain RMSD with respect to the crystal structure (Figure 3). In particular all seven final structures show that the TRP43 and/or TYR45 side chains point away from the α-helix, preventing the correct folding of the β3-β4 sheet, as shown in Figure 5.

These results are in agreement with experimental data,[62−64] indicating that the native-state fluorescence intensity of TRP43 is recovered more rapidly than the formation of stable hydrogen bonds in the β-sheets, thus implying that rapid partial or complete formation of the tertiary contacts between the β3-β4 sheet and the α-helix occurs.

To further verify this hypothesis two different strategies were used. In the first one the purpose was to obtain additional unfolded conformations with some native-like structural properties for the residues corresponding to the

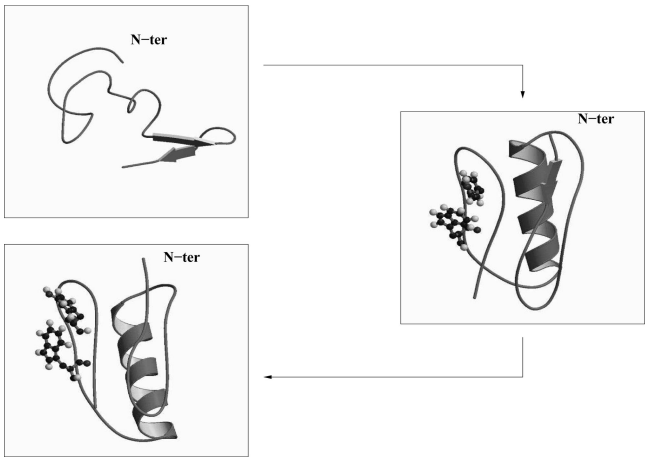

Figure 5. GB1. Backbone conformations at $t = 0$ ps, 746 ps, 4000 ps for RUN10. The side-chain orientations of TRP43 and TYR45 are also reported at $t = 746$ ps and $t = 4000$ ps.

β3 strand (residues 41−47). To this end a further unfolding simulation was performed starting from the structure extracted at $t = 4000$ ps of the native-state simulation and coupling the residues of the β3 strand to a thermal bath at $T = 300$ K, while the rest of the system was kept at $t = 500$ K. Two unfolded structures were extracted from this simulation and used as starting structures in the folding process (RUN11−12). Their conformations are shown in Figure 1. The structural properties of the starting and final conformations are reported in Table 1. A good agreement with the values obtained in the native-state simulation can be observed, thus indicating that the correct fold is obtained.

The second strategy consisted of using a new set of eigenvectors for the EDS simulations which included the Cα atoms of the whole protein together with the side-chain atoms of residues 41−47. Then, new simulations starting from the initial structures of the unsuccessful RUN4 and RUN8 using these new eigenvectors were performed. The results reported in Table 1 (RUN4′ and RUN8′) show that a correct folding was obtained in both cases.

3.3. SH3. The main structural properties of the native-state MD simulation at 300 K (NatSH3) are reported in Table 2. The data show a good agreement with the NMR structure. In the table are also reported the RMSD with respect to the NMR structure, Rg, N_β, and ρ values of the six unfolded structures used as starting points for the folding simulations. Figure 6 shows the corresponding structures together with the NMR one.

For each starting structure, the folding process was simulated for a time range of 3000−5000 ps by the EDS performed in a subspace defined by the last 160 eigenvectors over a total of 168 obtained from the native-state simulation (see the 'EDS Procedure' section). The final structural properties, averaged on the last 100 ps of each folding simulation, are reported in Table 2. The correct conformation was reached in five simulations (RUN1−5), that can be considered representative of the folding process, whereas the structure obtained in the sixth one showed values of RMSD, ρ, and N_β not in agreement with the values obtained in the native simulation.

The side-chain RMSD with respect to the NMR structure, averaged on the last 100 ps of the five EDS trajectories providing the correctly folded structures, is reported in Figure 7 and compared with the RMSD in the native-state simulation averaged on the equilibrated part of the simulation (200−5000 ps). It results in a good agreement between the two curves with the exception of the side chain of THR42 (a residue forming a β-turn), that shows a larger RMSD value at the end of the folding simulations than in the native one.

The analysis of the native contacts as a function of time shows that in four simulations, out of five correctly folded, the native interactions within the central β-sheet, consisting of the β2, β3, and β4 strands, precede the ones of the terminal β-sheet, consisting of the β1 and β5 strands, in agreement with experimental[47,48] and computational[43,46] data. As an example, the native contact maps, calculated at different times along the RUN3 trajectory, are reported in Figure 8 and compared with the contact map obtained from the native-state trajectory. The maps were calculated averaging over 10 ps starting at time $t = 0$, 100, 200, 2800 ps. A native contact between non-neighboring residues was considered to be formed if at least one distance between any two atoms was smaller than 0.6 nm. In the starting structure ($t = 0$ ps in Figure 8) the β2-β3 interaction is partially present; however, it is not complete, and the secondary structure is only partially formed (see structure RUN3 in Figure 6). At $t = 100$ ps the β2-β3 interactions are completely formed as well as the secondary structure, and part of the β1-β5 and the β3-β4 contacts are present. At $t = 200$ ps the β1-β5 and the β3-β4 interactions are almost completely formed, and at $t = 2800$ ps the contact map is similar to the one calculated on the native-state trajectory.

4. Conclusions

In the present work the EDS method is used to fold two small proteins, the all-β SH3 and the α/β GB1 protein, to their native structures starting from unfolded conformations and to reveal the known folding steps. The idea of the method is to bias the system toward its known native structure by means of a MD simulation, using a least biased procedure. This is accomplished by restraining only a subset of the degrees of freedom of the protein and by choosing such coordinates so to contain dynamical information of the native state. This is achieved by using a subgroup of the eigenvectors extracted from a principal component (or essential dynamics) analysis of the collective motions of the backbone Cα atoms of the protein in its native state. Hence, no information of the side chains is introduced. It is shown here that the EDS method does not "force", e.g., overcoming barriers higher than a few kT's, the simulation toward the correct folded structure; in fact not all the folding simulations were successful, in particular for the GB1 protein. When the protein gets into a nonproductive folding trap, the folded structure is not reached. Moreover, since the reaction coordinates used here contain information on the native state, it is possible with this procedure to find out the main mechanical information necessary for the folding process.

The results showed that in SH3 the native interactions within the central β-sheet precede the ones of the terminal

Table 2. SH3: Starting and Final Radius of Gyration (Rg_i, Rg_f), Backbone Root-Mean-Square Deviation ($RMSD_i$, $RMSD_f$) and Native Contact Fraction (ρ_i, ρ_f) with Respect to the Crystal Structure, Number of Residues in β-Structure ($N_{\beta i}$, $N_{\beta f}$) in the Folding Simulations[a]

SIM	R_{gi} (nm)	R_{gf} (nm)	$RMSD_i$ (nm)	$RMSD_f$ (nm)	ρ_i	ρ_f	N_{β_i}	N_{β_f}
NMR	1.04		–		–		21	
NatSH3	1.01(0.01)		0.19(0.03)		0.89(0.02)		22(2)	
RUN1	1.49	1.03	1.24	0.11	0.35	0.88	10	19
RUN2	1.29	1.03	1.05	0.19	0.27	0.86	11	20
RUN3	1.29	1.02	0.94	0.17	0.30	0.88	11	21
RUN4	1.33	1.03	1.22	0.10	0.21	0.85	0	20
RUN5	1.35	1.02	0.99	0.13	0.36	0.83	10	20
RUN6	1.34	1.08	1.35	0.47	0.16	0.50	5	7

[a] The final values in the folding simulations are averaged over the last 100 ps of each simulation. The values for the native-state trajectory (NatSH3) are averaged on the equilibrated part (200−5000 ps) with standard deviations in parentheses. In all folding simulations (RUN1-RUN6) the last 160 Cα eigenvectors are used in the EDS procedure.

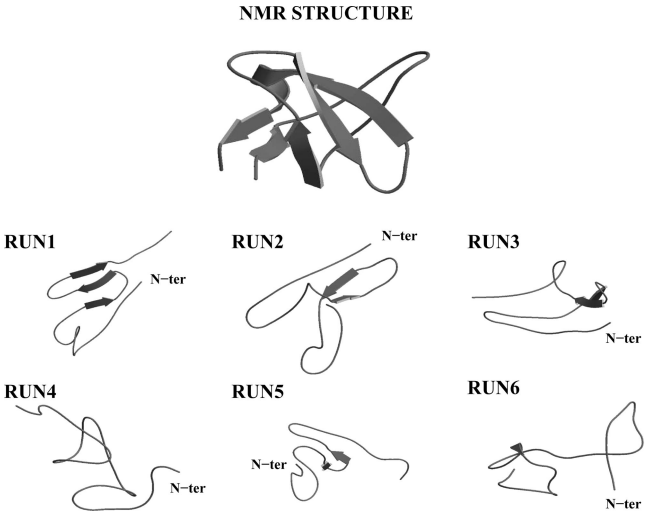

Figure 6. SH3. Backbone conformation of the NMR structure and of the six structures used as starting structures in the EDS folding simulations.

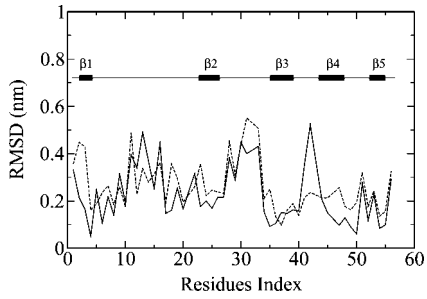

Figure 7. SH3. Side-chain RMSD with respect to the NMR structure. Solid line: average over the last 100 ps of the five correctly folded simulations (RUN1−5). Dashed line: average on the equilibrated part of the native-state simulation (200−5000 ps).

β-sheet, in agreement with experimental[47,48] and computational[43,46] data. In GB1, a correct folding of the side chains of TRP43 and TYR45 is a prerequisite for a correct folding, in agreement with experimental data[62−64] that show that the native-state fluorescence intensity of TRP43 is recovered more rapidly than the formation of stable β-sheet hydrogen bonds. These results, together with those previously reported for cytochrome *c*, confirm that EDS can detect the main structural characteristics of the folding mechanism. In this sense it could be used to predict crucial interactions in the folding of proteins, although validation from experiments is required.

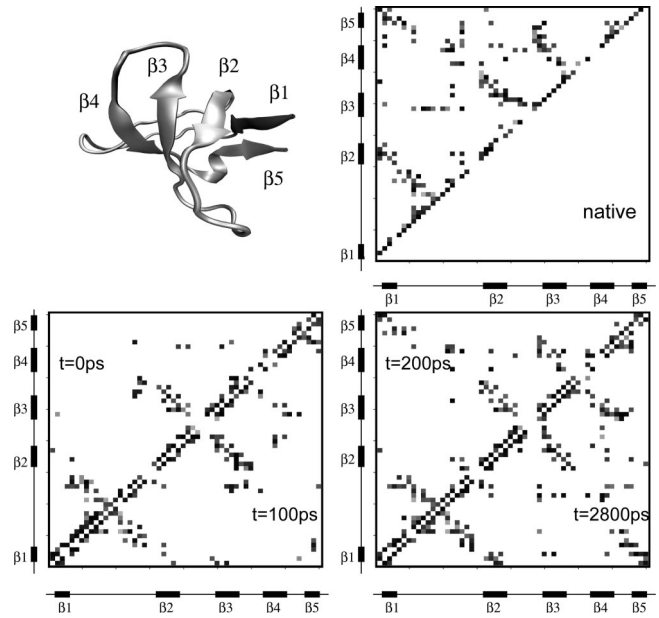

Figure 8. SH3. Native-contact maps. (top left side) Backbone conformation of the NMR structure showing the five β-strands: β1 residues 2−4, β2 residues 23−26, β3 residues 35−39, β4 residues 44−48, β5 residues 53−55. (top right side) Native-contact map obtained from the native-state trajectory. (Bottom) Native-contact maps at different times along RUN3, each calculated by averaging over 10 ps starting at $t = 0$ ps, 100 ps, 200 ps, and 2800 ps.

Acknowledgment. This work was supported by the Italian FIRB RBIN04PWNC_001 "Structure, function, dynamics and folding of proteins" founded by MIUR. We also acknowledge the University of Rome 'La Sapienza' for financial support with the project "MORFOGENESI MOLECOLARE: un approccio multidisciplinare per lo studio del folding e misfolding delle proteine" and CASPUR (Consorzio interuniversitario per le Applicazioni di Supercalcolo Per Universitá e Ricerca) for the use of its computational facilities.

References

(1) McCammon, J. A.; Gelin, B.; Karplus, M. *Nature* **1977**, *267*, 585–590.

(2) Wong, C. F.; Zheng, C.; Shen, J.; McCammon, J. A.; Wolynes, P. G. *J. Phys. Chem.* **1993**, *97*, 3100–3110.

(3) Dill, K.; Chan, H. *Nat. Struct. Biol.* **1997**, *4*, 10–19.

(4) Dobson, C. M.; Karplus, M. *Curr. Opin. Struct. Biol.* **1999**, *9*, 92–101.

(5) Onuchic, J. N.; Nymeyer, H.; García, A. E.; Chahine, J.; Socci, N. D. *Adv. Protein Chem.* **2000**, *53*, 87–152.

(6) Garcia-Mira, M. M.; Sadqi, M.; Fischer, N.; Sanchez-Ruiz, J. M.; Muñoz, V. *Science* **2002**, *298*, 2191–2195.

(7) Ulmschneider, J. P.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2004**, *126*, 1849–1857.

(8) Kubelka, J.; Hofrichter, J.; Eaton, W. A. *Curr. Opin. Struct. Biol.* **2004**, *14*, 76–88.

(9) Daura, X.; Gademann, K.; Juan, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. *Angew. Chem., Int. Ed.* **1999**, *38*, 236–240.

(10) Zhou, R.; Berne, B. J. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12777–12782.

(11) Daidone, I.; D'Abramo, M.; Di Nola, A.; Amadei, A. *J. Am. Chem. Soc.* **2005**, *127*, 14825–14832.

(12) Perez, A.; Luque, F. J.; Orozco, M. *J. Am. Chem. Soc.* **2007**, *129*, 14739–14745.

(13) Daidone, I.; Amadei, A.; Di Nola, A. *Proteins* **2005**, *59*, 510–518.

(14) Daidone, I.; Ulmschneider, M. B.; Di Nola, A.; Amadei, A.; Smith, J. C. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 15230–15235.

(15) Tirado-Rives, J.; Jorgensen, W. L. *Biochemistry* **1991**, *30*, 3864–3861.

(16) Tirado-Rives, J.; Jorgensen, W. L. *Biochemistry* **1993**, *32*, 4175–4184.

(17) Caflish, A.; Karplus, M. *J. Mol. Biol.* **1995**, *252*, 672–708.

(18) Lazaridis, T.; Lee, I.; Karplus, M. *Protein Sci.* **1997**, *6*, 2589–2605.

(19) Li, A. J.; Daggett, V. *J. Mol. Biol.* **1998**, *275*, 677–694.

(20) Mayor, U.; Johnson, C. M.; Daggett, V.; Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 13518–13522.

(21) Alonso, D. O. V.; Daggett, V. *Proc. Natl. Acad. Sci U.S.A.* **2000**, *97*, 133–138.

(22) Pan, Y.; Daggett, V. *Biochemistry* **2001**, *40*, 2723–2731.

(23) Sheinerman, F. B.; Brooks, C. L., III *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 1562–1567.

(24) Sheinermann, F. B.; Brooks, C. L., III *J. Mol. Biol.* **1998**, *278*, 439–456.

(25) Marchi, M.; Ballone, P. *J. Chem. Phys.* **1999**, *110*, 3697–3702.

(26) Paci, E.; Karplus, M. *J. Mol. Biol.* **1999**, *288*, 441–459.

(27) Paci, E.; Smith, L. J.; Dobson, C. M.; Karplus, M. *J. Mol. Biol.* **2001**, *306*, 329–347.

(28) Finkelstein, A. V. *Protein Eng.* **1997**, *10*, 843–845.

(29) Wang, T.; Wade, R. C. *J. Chem. Theory Comput.* **2007**, *3*, 1476–1483.

(30) Ferrara, P.; Apostolakis, J.; Caflisch, A. *Proteins* **2000**, *39*, 252–260.

(31) Gutin, A. M.; Abkevich, V. I.; Shakhnovich, E. I. *Fold. Des.* **1998**, *3*, 183–194.

(32) Klimov, D. K.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 2544–2549.

(33) Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D. *J. Mol. Biol.* **1997**, *268*, 209–225.

(34) Simons, K. T.; Ruczinski, I.; Kooperberg, C.; Fox, B.; Bystroff, C.; Baker, D. *Proteins* **1999**, *34*, 82–95.

(35) Bonneau, R.; Strauss, C. E.; Rohl, C. A.; Chivian, D.; Bradley, P.; Malmstrom, L.; Robertson, T.; Baker, D. *J. Mol. Biol.* **2002**, *322*, 65–78.

(36) Amadei, A.; Linssen, A. B. M.; de Groot, B. L.; van Aalten, D. M.; Berendsen, H. J. C. *J. Biomol. Struct. Dyn.* **1996**, *13*, 615–625.

(37) de Groot, B. L.; Amadei, A.; van Aalten, D. M. F.; Berendsen, H. J. C. *J. Biomol. Struct. Dyn.* **1996**, *13*, 741–751.

(38) García, A. E. *Phys. Rev. Lett.* **1992**, *66*, 2696–2699.

(39) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. *Proteins: Struct., Funct., Genet.* **1993**, *17*, 412–425.

(40) Meyer, T.; Ferrer-Costa, C.; Perez, A.; Rueda, M.; Bidon-Chanal, A.; Luque, F. J.; Laughton, C. A.; Orozco, M. *J. Chem. Theory Comput.* **2006**, *2*, 251–258.

(41) Harvey, S. C.; Gabb, H. A. *Biopolymers* **1993**, *13*, 741–751.

(42) Daidone, I.; Amadei, A.; Roccatano, D.; Di Nola, A. *Biophys. J.* **2003**, *85*, 2865–2871.

(43) Shea, J. E.; Onuchic, J. N.; Brooks, C. L., III *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 16064–16068.

(44) Guo, W.; Lampoudi, S.; Shea, J. E. *Biophys. J.* **2003**, *85*, 61–69.

(45) Guo, W.; Lampoudi, S.; Shea, J. E. *Proteins* **2004**, *55*, 395–406.

(46) Ding, F.; Guo, W.; Dokholyan, N. V.; Shakhnovich, E. I.; Shea, J. E. *J. Mol. Biol.* **2005**, *350*, 1035–1050.

(47) Riddle, D. S.; Grantcharova, P. V.; Santiago, J. V.; Alm, E.; Ruczinski, I.; Baker, D. *Nat. Struct. Biol.* **1999**, *6*, 1016–1024.

(48) Grantcharova, P. V.; Riddle, D. S.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *13*, 7084–7089.

(49) McCallister, L. E.; Alm, E.; Baker, D. *Nat. Struct. Biol.* **2000**, *7*, 669–673.

(50) Nauli, S.; Kuhlman, B.; Baker, D. *Nat. Struct. Biol.* **2001**, *8*, 602–605.

(51) van Gunsteren, W. F.; Berendsen, H. J. C. *Gromos manual*; BIOMOS, Biomolecular Software, Laboratory of Physical Chemistry, University of Groningen: The Netherlands, 1987.

(52) van Buuren, A. R.; Marrink, S. J.; Berendsen, H. J. C. *J. Phys. Chem.* **1993**, *97*, 9206–9212.

(53) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. *Intermolecular Forces*; Pullman, B., Ed.; D. Reidel Publishing Company: Dordrecht, The Netherlands, 1981.

(54) Ryckaert, J. P.; Bellemans, A. *Chem. Phys. Lett.* **1975**, *30*, 123–125.

(55) Brown, D.; Clarke, J. H. R. *Mol. Phys.* **1984**, *51*, 1243–1252.

(56) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(57) Yu, H.; Rosen, M. K.; Schreiber, S. L. *FEBS Lett.* **1993**, *324*, 87–92.

(58) Gallagher, T.; Alexander, P.; Bryan, P.; Gilliland, G. L. *Biochemistry* **1994**, *33*, 4721–4720.

(59) Roccatano, D.; Daidone, I.; Ceruso, M.-A.; Bossa, C.; Di Nola, A. *Biophys. J.* **2003**, *84*, 1876–1883.

(60) Shortle, D.; Ackerman, M. S. *Science* **2001**, *293*, 487–489.

(61) Zagrovic, B.; Snow, C. D.; Khaliq, S.; Shirts, M. R.; Pande, V. S. *J. Mol. Biol.* **2002**, *323*, 153–164.

(62) Park, S.; O'Neil, K. T.; Roder, H. *Biochemistry* **1997**, *36*, 14277–14283.

(63) Park, S.; Ramachandra Shastry, M. C.; Roder, H. *Nat. Struct. Biol.* **1999**, *6*, 943–947.

(64) Kuszewski, J.; Clore, G. M.; Gronenborn, A. M. *Protein Sci.* **1994**, *3*, 1945–1952.

CT800157V