JCTC Journal of Chemical Theory and Computation

A Collective Variable for the Efficient Exploration of Protein Beta-Sheet Structures: Application to SH3 and GB1

Fabio Pietrucci* and Alessandro Laio

International School for Advanced Studies (SISSA-ISAS), via Beirut 2-4, I-34014 Trieste, Italy

Received April 24, 2009

Abstract: We introduce a new class of collective variables which allow forming efficiently beta-sheet structures in all-atom explicit-solvent simulations of proteins. By this approach we are able to systematically fold a 16-residue beta hairpin using metadynamics on a single replica. Application to the 56-residue SH3 and GB1 proteins show that, starting from extended states, in ∼100 ns tens of structures containing more than 30% beta-sheet are obtained, including parts of the native fold. Using these variables may allow folding moderate size proteins with an accurate explicit solvent description. Moreover, it may allow investigating the presence of misfolded states that are relevant for diseases (e.g., prion and Alzheimer) and studying beta-aggregation (amyloid diseases).

1. Introduction

All-atom explicit solvent simulations are still not competitive with bioinformatics or knowledge-based approaches for protein structure predictions, at least in terms of cost-effectiveness. In order to improve the predictivity of accurate simulations, it will be necessary to address two issues, both of the utmost importance: First, the accuracy of the force fields that, as is well-known, is still far from optimal. Second, the sampling efficiency: even if the "perfect" force field was available, in order to predict the native fold of a protein one should explore several structures and compare their free energies accurately. In this work we address the problem of performing an efficient sampling of protein conformations.

The structure of proteins typically contains a large amount of secondary structure, in the form of alpha helices and beta sheets. The formation of an α helix is a rather simple process, which mainly requires the local alignment of backbone dihedrals in a segment of the protein chain and typically happens on a time scale of about 100 ns.[1] Instead, beta-structures are more complicated, as shown by characteristic formation times at least 1 order of magnitude longer.[1−3] This difference is mainly related to the fact that building of beta-structures requires the proper dihedral arrangement in two distant segments of the protein chain and the simultaneous formation of specific interstrand H-bonds. As a result, simulating by accurate molecular dynamics (MD) with explicit solvent the folding process of proteins containing beta-structure is challenging: already studying a 16-residues beta-hairpin requires significant computational resources.[4−6] This is an important limitation, as beta-structures are present in many proteins and, moreover, are the key structural element in fibrils.[7]

In order to enhance the probability of observing beta-structures one can use an enhanced sampling scheme such as umbrella sampling, thermodynamic integration, or metadynamics.[8] These approaches require choosing an appropriate collective variable (CV) which describes the progress of the conformational transition. For instance, such a variable could be defined using the beta secondary structure definition of DSSP[9] or STRIDE,[10] which is primarily based on the H-bonds pattern. Unfortunately, it is well-known that a precise indicator of a structural property is not necessarily a good *reaction coordinate* for simulating a transition process.[11] Indeed, we tested CVs based on the peculiar H-bond arrangement of beta-structures, that, for example, in the antiparallel β sheet are formed between pairs of residues $(i,i+h)$, $(i+2,i+h-2)$, etc. We observed that using this class of CVs with metadynamics[12] allows the formation of beta-structure but is affected by technical problems when implemented in MD, since it drives the system not only toward well-formed beta strands but also toward unphysical structures with unlikely conformations. This can be understood considering that the formation of even a single beta bridge is a rather complex process that requires establishing selected hydrogen bonds after specific dihedral transitions and after the correct alignment of the two chain segments forming the bridge. A CV taking into account only one of these three aspects may not be effective to bias the formation of beta-sheet structures.

To overcome these limitations, we here introduce a CV for beta-structure which is defined in a different manner, by counting how many pairs of 3-residue segments adopt the correct beta conformation in a given protein structure. The correct conformation is taken simply as the average beta conformation of experimental protein structures. The CV is not tailored on the specific fold of a single protein, but it is meant as a general-purpose CV which can describe beta-structure in all proteins.

* Corresponding author e-mail: fabio.pietrucci@gmail.com.

10.1021/ct900202f CCC: $40.75
Published on Web 08/04/2009

First, we benchmarked the new CV by simulating the C-terminal beta hairpin of protein GB1 in explicit solvent, using a single-replica metadynamics[12,13] simulation. The correct folded state is systematically obtained, together with several misfolded states, at a cheap computational cost. Next, we investigated the conformational space of two larger (56 amino acids) proteins, SH3 and GB1, whose native folds include a large amount of beta secondary structure. These two proteins have been investigated in several experimental and theoretical studies of protein folding.[14−18] However, a compelling simulation of their folding mechanism with accurate explicit-solvent force fields is still lacking. We attempted to fold both SH3 and GB1 by bias-exchange metadynamics[19,20] simulations in explicit solvent, employing the new CVs. Starting from extended states, within ∼100 ns of total simulated time tens of different conformations with a large content of extended beta structure are obtained. Among these, several configurations are obtained which contain the main structural elements of the native state. Bias-exchange metadynamics simulations performed on the same systems but using other variables explore the conformational space ∼10 times less efficiently.

2. The Beta Collective Variable

In order to define a CV for exploring protein beta-sheet structures by metadynamics or other enhanced sampling techniques, we first defined from the PDB database the shape of an ideal building block for beta sheets, i.e. a small beta subunit composed of a few amino acids, which by replication gives rise to the extended beta-strucures. To this aim, we considered the representative proteins of the 20 architecture entries in the "mainly beta" class of the CATH database[21] (PDB codes 1bds, 1gvk, 1h8p, 1i5p, 1itv, 1k7i, 1m3y, 1n7v, 1nh2, 1qre, 1rg8, 1tl2, 1w6s, 1ylh, 2bbk, 2dpf, 2hnu, 2nwf, 3sil, 4bcl). For each protein, we extracted the residues belonging to beta secondary structure (according to the STRIDE[10] definition), and, among them, we extracted all pairs of segments of 3 residues connected by hydrogen bonds. We computed the RMSD of the positions of backbone N, C_α, C, O, and C_β atoms in the 3 + 3 blocks. Antiparallel beta blocks are similar within a RMSD of only 0.048 ± 0.017 nm, parallel beta blocks within 0.066 ± 0.028 nm. Therefore the beta-structures observed in proteins, despite their impressive variety, are composed of 3 + 3 blocks which are remarkably similar. This allows the definition of the ideal (i.e., average) "beta block". We defined the ideal antiparallel and parallel beta blocks by taking the central structure of each pool. In the case of the parallel beta-structure, two equivalent blocks exist, corresponding to a symmetry operation obtained by rotating of 180° both the 3-residue segments around their backbone axis.

Using this definition of beta blocks, we implemented a CV which counts how many 3 + 3 residues units are similar to the "beta block". This CV is defined as a differentiable function of the atomic coordinates in the following manner

$$S = \sum_{\alpha} n[\mathrm{RMSD}(\{\mathbf{R}_i\}_{i\in\Omega_\alpha}, \{\mathbf{R}^0\})] \qquad (1)$$

$$n(\mathrm{RMSD}) = \frac{1 - (\mathrm{RMSD}/0.1)^8}{1 - (\mathrm{RMSD}/0.1)^{12}} \qquad (2)$$

where n is a function switching smoothly between 0 and 1, the RMSD is measured in nm, and $\{\mathbf{R}_i\}_{i\in\Omega_\alpha}$ are the atomic coordinates of a set Ω_α of six residues of the protein, while $\{\mathbf{R}^0\}$ are the corresponding atomic positions of the ideal beta block. In the case of antiparallel beta, all sets Ω_α of residues of the form $(i,i+1,i+2; i+h+2,i+h+1,i+h)$ are summed over in eq 1. For parallel beta, sets $(i,i+1,i+2;i+h,i+h+1,i+h+2)$ are instead considered. For each residue, only backbone N, C_α, C, O, and C_β atoms are included in the RMSD calculation (in Gly residues the C_β is missing and the corresponding hydrogen is used instead).

The same procedure has been applied to define the ideal α helix block formed by six consecutive residues, in order to define a CV measuring the amount of alpha secondary structure. In this case the sum in eq 1 runs over all possible sets Ω_α of six consecutive protein residues $(i, i+1,i+2,i+3,i+4,i+5)$, and $\{\mathbf{R}^0\}$ are the atomic positions of the ideal alpha block. In summary, the CVs $S_{\mathrm{anti}\beta}$, $S_{\mathrm{para}\beta}$, and S_α are approximately proportional to the number of beta/alpha blocks of six residues which are present in a protein 3D structure (Figure 1).

3. Metadynamics Simulation of Beta-Hairpin Folding

We tested on several proteins the ability of the new CVs to generate secondary structure, starting from unfolded conformations and performing metadynamics simulations[12] in explicit solvent. First, we benchmarked our approach on the 16-residues C-terminal beta hairpin of protein GB1 (PDB code 1pgb, Figure 2-A). We used a version of the GROMACS 3.3.1 package[22] modified by us, employing the AMBER03[23] and TIP3P[24] force fields for protein and water, respectively. The protein was solvated by 3373 water molecules in a orthorhombic box of 105.8 nm^3, neutralized by three Na^+ ions. The particle-mesh Ewald method[25,26] was used for long-range electrostatics with a short-range cutoff of 0.8 nm. A cutoff of 0.8 nm was used for the Lennard-Jones interactions. All bond lengths were constrained to their equilibrium length with the LINCS[27] algorithm. The time step for the MD simulation was 2.0 fs. NPT simulations at 340 K and 1 atm were performed by coupling the system to a Nose-Hoover thermostat[28,29] and a Berendsen barostat,[30] both with relaxation time of 1 ps. After 1 ns of

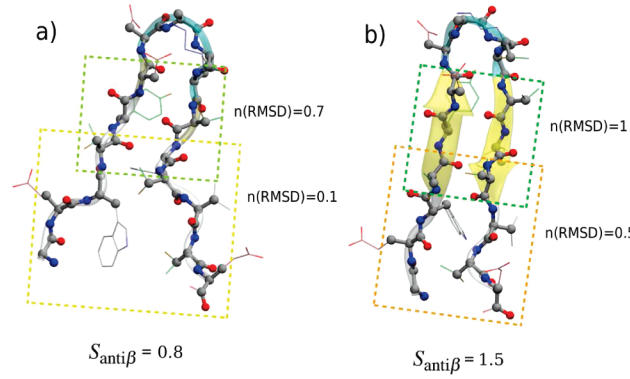

Figure 1. Values assumed by the CV $S_{\mathrm{anti}\beta}$ in the (a) partially formed and (b) almost-completely folded beta hairpin. The dashed rectangles outline 3 + 3 residue beta blocks.

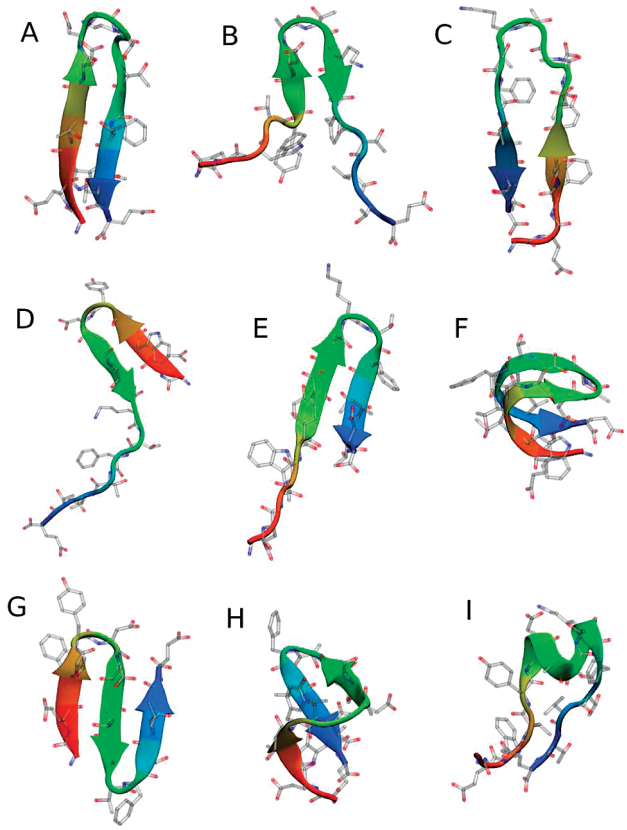

Figure 2. Folded state (A) and several misfolded conformations (B−I) of the 16-residues C-terminal beta-hairpin of GB1, obtained by a single metadynamics simulation of 100 ns. The collective variables $S_{anti\beta}$ and C_α radius of gyration have been biased.

equilibration, the barostat was removed, and the simulations were continued in the NVT ensemble.

Four independent metadynamics simulations of length 100 ns were performed, starting from extended states without secondary content. The CVs antiparallel beta ($S_{anti\beta}$) and radius of gyration of C_α (R_{gyr}) were biased by 2-dimensional Gaussians of height 2 kJ/mol, width 0.1 and 0.05 nm, respectively, adding a Gaussian every 5 ps.

In each simulation, the experimental folded state is observed within 50 ns (backbone RMSD < 0.2 nm, Figure 2-A). Furthermore, several different misfolded states with alpha or beta content are also found, some of which are reported in Figure 2 (panels B−I). This calculation shows that a suitable choice of the reaction coordinate allows folding the beta hairpin and finding misfolded states even using a single replica.

4. Bias-Exchange Metadynamics Simulation of SH3 Folding

Using the CV introduced in this work, we also investigated the conformational space of the larger (56 amino acids) protein src-SH3. The native fold consists of a terminal beta-hairpin packed orthogonally on top of a three-stranded antiparallel beta-sheet, plus a small 3_{10} helix (PDB codes 1srl, Figure 3-A). The protein was solvated by 3641 water molecules in a cubic box of 127.2 nm^3, neutralized by three Na^+ ions. The MD parameters are the same as for the beta hairpin (see above).

We performed a bias-exchange metadynamics[19] simulation at 340 K, employing four replicas and starting from an extended

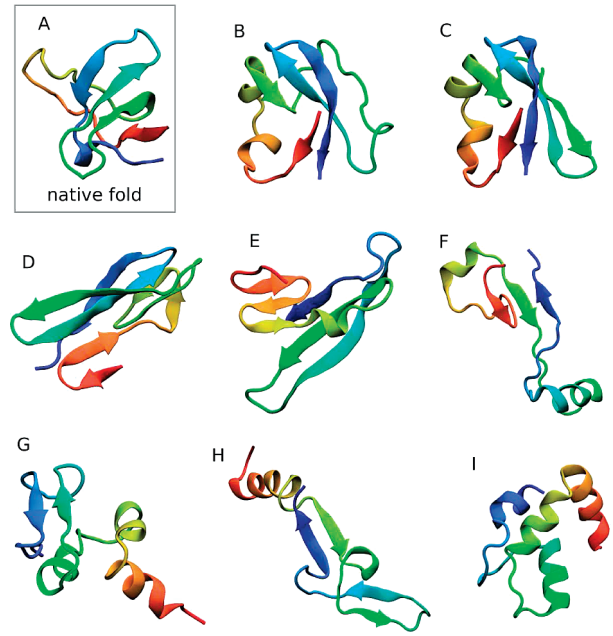

Figure 3. Experimental folded state of SH3 (A) and selected conformations obtained using the CVs introduced in this work in a bias-exchange metadynamics simulation (B−I).

Table 1. Average Number of SH3 Clusters Explored per 100 ns of Total Simulation Time by Bias-Exchange Metadynamics, Using the New CVs or the Old CVs[a]

	old CVs	new CVs
>20% sec. str.	97	203
>30% sec. str.	42	104
>10% beta	4	40
>20% beta	1.7	15
>30% beta	0.1	4

[a] See the text for a definition. The total simulation time is 940 ns for the old CVs and 240 ns for the new CVs. The structures have been clustered using the TM-align algorithm[31] with a threshold of 0.5.

state with no secondary structure. The following CVs have been biased, each one on a different replica: S_α, $S_{anti\beta}$, $S_{para\beta}$, and the radius of gyration of hydrophobic side chain carbons (R_{gyr}). These CVs are not tailored on the specific native state of SH3, but they represent general-purpose reaction coordinates for protein folding. One-dimensional hills of height 2 kJ/mol were added every 5 ps, and exchanges of the bias potentials were attempted every 50 ps. The trajectories were clustered with the TM-align algorithm[31] (threshold 0.5) in order to extract the significantly different structures.

Within 60 ns per replica (total simulation time 240 ns), ~200 different structures with more than 30% content of secondary structure are obtained (see Table 1), a selection of which is reported in Figure 3. In particular, 9 different structures are obtained which include more than 30% extended beta. Among these, structures B and C in Figure 3 clearly contain a substantial part of the native multiple-β sheet of SH3 (Figure 3-A), although with some topological differences.

As a comparison, another bias-exchange metadynamics simulation was performed on the same system, this time without employing the new CVs introduced in this work but biasing the CVs introduced in ref 19: number of backbone H-bonds in the first half of the protein, in the second half, and between the

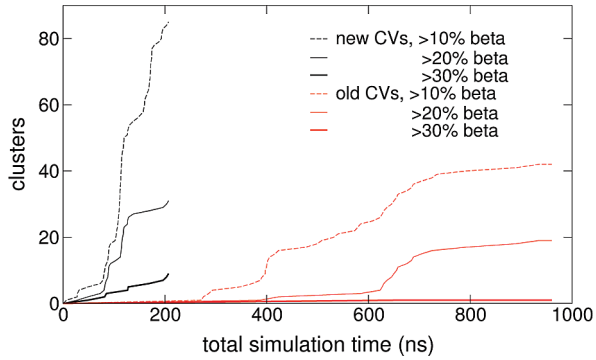

Figure 4. Number of SH3 clusters explored by enhanced sampling molecular dynamics as a function of total simulation time, using the new CVs or the old CVs (see text for definition). Clusterization has been performed using the TM-align algorithm[31] with a threshold of 0.5. The clusters contain an amount of beta-structure between 10% and 30%, following the DSSP definition.[9]

two halves; helicity of the backbone Φ_α (as defined in ref 19) in each fourth of the protein chain; number of contacts among aromatic side chain carbons, or among hydrophobic side chain carbons, defined as $N = \Sigma_{ij}[1 - (R_{ij}/R_0)^8]/[1 - (R_{ij}/R_0)^{14}]$ ($R_0 =$ 0.3 nm, i,j run over the appropriate carbon atoms). Each replica was biased by two-dimensional hills of height 0.5 kJ/mol added every 2 ps; exchanges of the bias potentials were attempted every 2 ps. These variables are referred to as "old CVs" in Table 1. Sixteen replicas have been used, and a simulation of 60 ns per replica was performed starting from extended states. By using these CVs a much smaller number of structures containing beta-sheets is explored per unit simulation-time, compared to the new CVs (Table 1 and Figure 4).

5. Bias-Exchange Metadynamics Simulation of GB1 Folding

As a second application, we studied the 56-residues protein GB1. The native fold consists of a four-strands β sheet formed by two terminal beta-hairpins connected in parallel and packed on top of an α helix (PDB code 1pgb, Figure 5-A). The protein was solvated by 3780 water molecules in a cubic box of 125.0 nm^3, neutralized by four Na^+ ions. The MD parameters are the same as above.

A bias-exchange metadynamics[19] simulation was performed on GB1 at 300 K, using 8 replicas and biasing each of the following CVs on a different replica: $S_{anti\beta}$, $S_{para\beta}$, the helicity of the backbone Φ_α (as defined in ref 19) in each third of the protein chain, and the number of contacts $N = \Sigma_{ij}[1 - (R_{ij}/R_0)^8]/[1 - (R_{ij}/R_0)^{14}]$ ($R_0 = 0.7$ nm) with the summation extended to C_α pairs belonging to first and second, second and third, or first and third segments of the protein chain. One-dimensional hills of height 2.5 kJ/mol were added every 5 ps, and exchanges of the bias potentials were attempted every 50 ps. The trajectories were clustered with the TM-align algorithm[31] as described above.

Starting from extended states, in 80 ns per replica 53 different structures with more than 30% beta content are explored. A selection of the structures is reported in Figure 5, panels B—Q. Remarkably, structures B, D, E, and P contain one or both of the native terminal beta hairpins, whereas structures B, C, D, E, and L contain the native central helix (compare Figure 5-A). Thus, also in this case, using the newly introduced variable allows exploring a very large number of structures with significant secondary content.

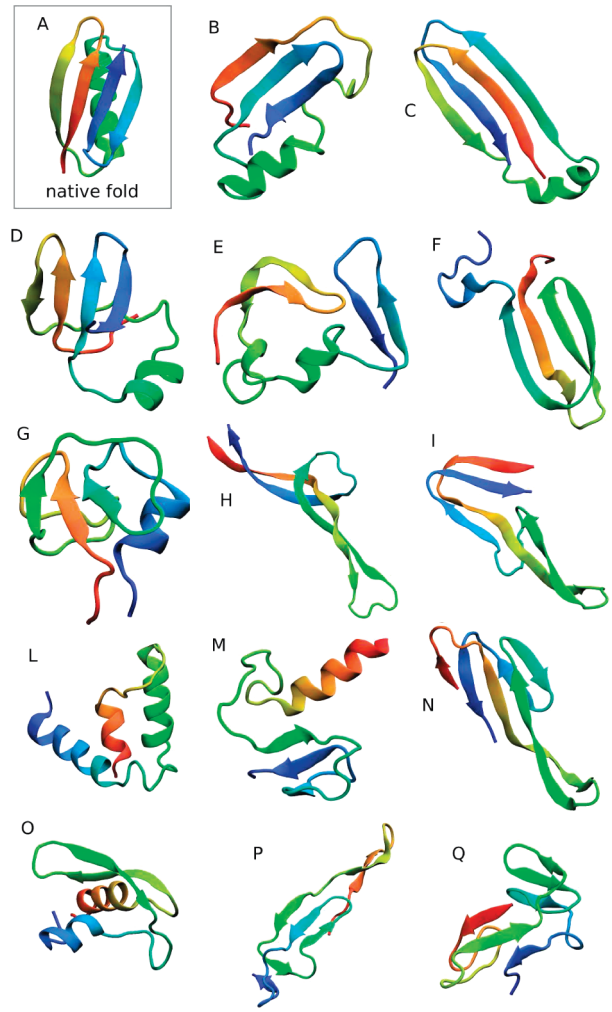

Figure 5. Experimental folded state of GB1 (A) and selected conformations obtained using the CVs introduced in this work in a bias-exchange metadynamics simulation (B—Q).

6. Conclusions

We introduced a new class of collective variables (CVs) specifically designed for observing the formation of beta sheets and alpha helices. The CVs are not tailored specifically on a given protein fold, but they are aimed at describing the content of secondary structure in all proteins. Using these CVs together with an enhanced sampling technique such as umbrella sampling, thermodynamic integration, or (bias-exchange) metadynamics allows exploring quickly a large number of complex alpha- and beta-structures starting from unfolded states and employing accurate explicit-solvent force fields. In particular, it is possible to systematically fold the C-terminal beta-hairpin of protein GB1 employing a single-replica simulation. Application of the new CVs to the 56-residue proteins SH3 and GB1 in explicit solvent shows that bias-exchange metadynamics simulations allow to observe in ~100 ns the formation of tens of different structures with large alpha- and beta-content. Some of these structures contain parts of the elements of the native

fold. For both proteins, however, the exact native fold is not reached within the relatively short span of the simulations.

One should remark that the experimental folding times of GB1 and SH3 are of the order of milliseconds and seconds, respectively, indicating that the atomic rearrangements that have to take place in order to explore the folded state are rather complex. For a comparison, the folding times of GB1 and SH3 are 3 to 6 orders of magnitude larger than the one of advillin, the largest system that has been so far reversibly and reproducibly folded with an all atom force field describing the solvent explicitly. The quality of the new variables introduced here is demonstrated only comparing the number of nontrivial structures that are found in a given simulation time. More extended simulations should be employed to find the experimental folded state and also to estimate the relative free energy of different conformations, e.g. by means of a weighted hystogram analysis of the bias-exchange metadynamics trajectories, as detailed in ref 32.

Still, our results allow for being optimistic about the possibility to completely fold proteins of less than 100 amino acids using simulation times of the order of microseconds. Furthermore, the new variables may help investigating the presence of protein misfolded states and the phenomenon of beta-aggregation, which are relevant to understand the mechanism of several diseases.

Acknowledgment. We acknowledge Pilar Cossio, Fabrizio Marinelli, and Xevi Biarnés for useful discussions. We also acknowledge the grant MIUR PRIN-2006025255 and CINECA for providing computational resources.

References

(1) Eaton, W. A.; Munoz, V.; Hagen, S. J.; Jas, G. S.; Lapidus, L. J.; Henry, E. R.; Hofrichter, J. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 327–359.

(2) Munoz, V.; Thompson, P. A.; Hofrichter, J.; Eaton, W. A. *Nature* **1997**, *390*, 196–199.

(3) Jager, M.; Deechongkit, S.; Koepf, E. K.; Nguyen, H.; Gao, J.; Powers, E. T.; Gruebele, M.; Kelly, J. W. *Biopolymers* **2008**, *90*, 751–758.

(4) Bolhuis, P. G. *Biophys. J.* **2005**, *88*, 50–61.

(5) Bussi, G.; Gervasio, F. L.; Laio, A.; Parrinello, M. *J. Am. Chem. Soc.* **2006**, *128*, 13435–13441.

(6) Yoda, T.; Sugita, Y.; Okamoto, Y. *Proteins* **2007**, *66*, 846–859.

(7) Chiti, F.; Dobson, C. M. *Annu. Rev. Biochem.* **2006**, *75*, 333–366.

(8) Dellago, C.; Bolhuis, P. G. *Advanced Computer Simulation Approaches for Soft Matter Sciences III*; Springer: Berlin/Heidelberg, 2008; Vol. 221, pp 1−67.

(9) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577–2637.

(10) Frishman, D.; Argos, P. *Proteins* **1995**, *23*, 566–579.

(11) Geissler, P. L.; Dellago, C.; Chandler, D. *J. Phys. Chem. B* **1999**, *103*, 3706–3710.

(12) Laio, A.; Gervasio, F. L. *Rep. Prog. Phys.* **2008**, *71*, 126601.

(13) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562–12566.

(14) Shea, J. E.; Brooks, C. L. *Annu. Rev. Phys. Chem.* **2001**, *52*, 499–535.

(15) Hubner, I. A.; Edmonds, K. A.; Shakhnovich, E. I. *J. Mol. Biol.* **2005**, *349*, 424–434.

(16) He, Y.; Chen, Y.; Alexander, P.; Bryan, P. N.; Orban, J. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 14412–14417.

(17) Narzi, D.; Daidone, I.; Amadei, A.; Di Nola, A. *J. Chem. Theory Comput.* **2008**, *4*, 1940–1948.

(18) Hori, N.; Chikenji, G.; Berry, R. S.; Takada, S. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 73–78.

(19) Piana, S.; Laio, A. *J. Phys. Chem. B* **2007**, *111*, 4553–4559.

(20) Piana, S.; Laio, A.; Marinelli, F.; Troys, M. V.; Bourry, D.; Ampe, C.; Martins, J. C. *J. Mol. Biol.* **2008**, *375*, 460–470.

(21) Cuff, A. L.; Sillitoe, I.; Lewis, T.; Redfern, O. C.; Garratt, R.; Thornton, J.; Orengo, C. A. *Nucleic Acids Res.* **2009**, *37*, D310–D314.

(22) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Model.* **2001**, *7*, 306–317.

(23) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J. M.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999–2012.

(24) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

(25) Darden, T. A.; York, D. *J. Chem. Phys.* **1993**, *98*, 10089.

(26) Essman, U.; Perera, L.; Berkowitz, M. L.; Darden, T. A.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577.

(27) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, G. E. M. J. *J. Comput. Chem.* **1997**, *18*, 1463.

(28) Nose, S. *Mol. Phys.* **1984**, *52*, 255.

(29) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695.

(30) Berendsen, H. J. C.; Postma, J. P. M.; Gusteren, W. F. V.; Nola, A. D.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.

(31) Zhang, Y.; Skolnick, J. *Nucleic Acids Res.* **2005**, *33*, 2302–2309.

(32) Marinelli, F.; Pietrucci, F.; Laio, A.; Piana, S. *PLoS Comput. Biol.* 2009. in press.

CT900202F