Folding of a miniprotein with mixed fold

Sandipan Mohanty, and U. H. E. Hansmann

Folding of a miniprotein with mixed fold

Sandipan Mohanty[a)]
John von Neumann Institut für Computing, Forschungszentrum Jülich, Jülich D-52425, Germany

U. H. E. Hansmann[b)]
Department of Physics, Michigan Technological University, Houghton, Michigan 49931-1295, USA
and John von Neumann Institut für Computing, Forschungszentrum Jülich, Jülich D-52425, Germany

Using the 28 residue $\beta\beta\alpha$ protein FSD-EY as a target system, we examine correction terms for the ECEPP/3 force field. We find an increased probability of formation of the native state at low temperatures resulting from a reduced propensity to form α helices and increased formation of β sheets. Our analysis of the observed folding events suggests that the C-terminal helix of FSD-EY is much more stable than the N-terminal β hairpin and forms first. The hydrophobic groups of the helix provide a template which promotes the formation of the β hairpin that is never observed to form without the helix. © *2007 American Institute of Physics.* [DOI: 10.1063/1.2753835]

I. INTRODUCTION

Most proteins fold into unique three-dimensional native states. This is because those states have the lowest free energy among all conformations the protein can take and hence are thermodynamically most probable. Simulating this process on a computer has proven extremely challenging. This is in part due to the difficulties in formulating an effective force field function that captures accurately the energetic and entropic balance within a solvated protein. The problems are amplified by the extremely large conformational space of proteins. For proteins large enough to have stable native states, it is rarely clear whether its conformational space was sufficiently searched and the minimum energy state was indeed found.

Very small peptides, for which a thorough search of the conformational space is feasible, seldom have stable native states. The few small proteins that have experimentally well characterized native states offer an important bridge between the competing needs of completeness of search and concreteness of experimental properties. In this article, we study the folding of one of these small proteins, FSD-EY (PDB ID: 1FME), that has a $\beta\beta\alpha$ structural motif[1] (cf. Fig. 1). Proteins with such mixed secondary structure content pose a specially great challenge, as most all-atom force fields tend to have a bias toward either helical or β-sheet secondary structures. For instance, the ECEPP/3 force field seems to put an overemphasis of helical structures. Abagyan and Totrov[2] have proposed two correction terms to the ECEPP potential that slightly destabilize helical structures. They tested the new terms by simulating the 23 residue $\beta\beta\alpha$ miniprotein BBA5 (PDB ID: 1T8J) which at the turn of the β hairpin has a non-natural amino acid, a D-proline. Here, we test this modification for a protein with only natural amino acids and study

the relative importance of the two correction terms. We find that the C-terminal helix of FSD-EY forms first and is more stable than the N-terminal β hairpin. The latter forms only after the hydrophobic groups of the helix provide a template.

II. MODEL AND FORCE FIELD CORRECTIONS

In this study, we have used the program package simple molecular mechanics for proteins SMMP.[3] SMMP implements a flexible representation of the polypeptide chains. In principle, all bond lengths and bond angles could be used as degrees of freedom. However, in the temperature ranges relevant in protein folding simulations, fluctuations of bond lengths and bond angles are much less important than the torsional degrees of freedom, and hence they are kept fixed, which leaves the backbone and side chain torsional degrees of freedom as the only remaining degrees of freedom in the system.

Several interaction potentials are implemented in SMMP. For this work, we use the ECEPP/3 force field,[4–7] a solvation term,[8] along with two small correction terms. The ECEPP/3 force field approximates the interactions between the atoms within a protein by the sum of electrostatic energy E_C, a Lennard-Jones term E_{LJ}, hydrogen-bonding term E_{HB}, and a torsion energy E_{tor},

$$E_{\text{ECEPP/3}} = E_C + E_{LJ} + E_{HB} + E_{tor} = \sum_{(i,j)} \frac{332 q_i q_j}{\epsilon r_{ij}}$$
$$+ \sum_{(i,j)} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + \sum_{(i,j)} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right)$$
$$+ \sum_{l} U_l (1 \pm \cos(n_l \xi_l)), \tag{1}$$

where r_{ij} is the distance between the atoms i and j, ξ_l is the lth torsion angle, and energies are measured in kcal/mol.

The protein-solvent interactions are approximated by a solvent accessible surface term,

a)Electronic mail: s.mohanty@fz-juelich.de
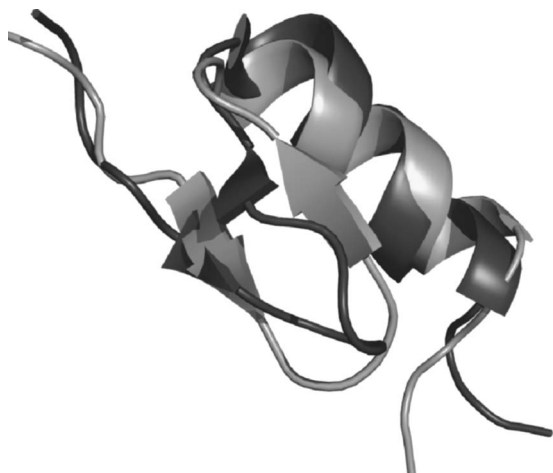b)Electronic mail: u.hansmann@fz-juelich.de and hansmann@mtu.edu

FIG. 1. Optimal overlap of a simulation structure from the minimum corresponding to the native structure and the experimental structure. Note that there are large fluctuations in both the turn and the strands of the β hairpin, so that the native minimum turns out to be rather broad in the simulations [rendered with PYMOL (Ref. 21)].

$$E_{\text{solv}} = \sum_i \sigma_i A_i. \qquad (2)$$

The sum goes over the solvent accessible areas A_i of all atoms i, weighted by solvation parameters σ_i, as determined in Ref. 8, a common choice when the ECEPP/3 force field is utilized. While this implicit solvent is a crude approximation of the protein-solvent interaction, combinations of fixed-geometry all-atom models with solvent accessible surface terms have recently proven surprisingly successful.[9–13]

In the case of the ECEPP/3 potential, previous experience suggests that helical structure elements are energetically overstabilized. In Ref. 2, two correction terms were proposed to suppress slightly the helical tendencies as well as to account for side chain entropies. The form of the corrected interaction potential used here is

$$E = E_{\text{ECEPP3}} + E_{\text{RC}} + E_{\text{ent}},$$

$$E_{\text{RC}} = \sum_i k_{\text{RC}}^{(i)}(1 - \sin(\psi_i)), \qquad (3)$$

$$E_{\text{ent}} = -RT\sum_i \alpha_i \Delta S_i,$$

where i in the summations runs over the residues, ψ_i is the backbone ψ angle of the ith residue, α_i is the ratio of the exposed area of the ith residue and the maximum exposed area for that side chain, and ΔS_i is the burial entropy, a term that approximately accounts for side chain entropy by connecting it to how exposed a side chain is to the solvent.

Originally, the term E_{ent} was developed as a quick estimate of side chain entropies. A completely buried side chain is assigned zero entropy and a fully exposed one a maximum entropy, while partially exposed side chains are assigned interpolated values. Within a canonical simulation, this is not a reasonable concept. However, by fixing $T=300$ K, E_{ent} be-

comes proportional to the solvent accessible surface area of the residue, i.e., can be interpreted as a correction term to E_{solv}.

The term E_{RC} corrects the intramolecular energy as described by the ECEPP force field by slightly destabilizing the α helical region of the Ramachandran plots. The strength of this term is 0.5 kcal/mol for all residue species except for valine, threonine, and isoleucine for which the strength is 1 kcal/mol. This term can be understood as an adjustment of the torsion energy term E_{tor} in ECEPP. The van der Waals interactions and the electrostatic interactions of the atoms separated by three covalent bonds contribute directly toward the observed torsional barriers. The parameters for the torsional energy term therefore depend on the other nonbonded interactions and have to be adjusted to match the dihedral angle distributions to those extracted from experiments or quantum chemistry calculations. As detailed quantum chemistry calculations are only feasible for very small systems, consisting of a few residues, the resulting parametrization of E_{tor} may lead to incorrect dihedral angle distribution and a bias for certain secondary structure elements. The E_{RC} term in Eq. (3) attempts to adjust empirically the torsional potential of backbone dihedral angles based on its effect on full length folding simulations. Note that in recent years the torsional energy terms in several commonly used force fields have undergone corrections to improve the description of secondary structure propensities in proteins.[14]

The competing interactions in this detailed energy function lead to an energy landscape that is characterized by a multitude of minima separated by high energy barriers. As the probability to cross an energy barrier of height ΔE is given by $\exp(-\Delta E/k_B T)$ (k_B the Boltzmann constant), it follows that extremely long runs are necessary to obtain sufficient statistics in regular canonical simulations at a low temperature T. One popular method to overcome the resulting extremely slow thermalization at low temperatures is parallel tempering[15,16] (also known as replica exchange method or multiple Markov chains), a technique first applied to protein studies in Ref. 17. In its most common form, one considers an artificial system built up of N *noninteracting* replicas of the molecule, each at a different temperature T_i. In addition to standard Monte Carlo or molecular dynamics moves that act only on one replica (i.e., the molecule at a fixed temperature), an exchange of conformations between two copies i and $j=i+1$ is allowed with probability

$$w(\mathbf{C}^{\text{old}} \rightarrow \mathbf{C}^{\text{new}}) = \min(1, \exp(-\beta_i E(C_j) - \beta_j E(C_i)$$
$$+ \beta_i E(C_i) + \beta_j E(C_j))). \qquad (4)$$

The exchange of conformations leads to a faster convergence of the Markov chain at low temperatures than is observed in regular canonical simulations with only local moves. The resulting random walk in temperatures allows the configurations to move out of local minima and cross energy barriers.

Besides replica exchange moves, configurations are updated both through simple rotations by a randomly chosen value of backbone and side chain torsional angles and through so-called biased Gaussian steps[18] (BGS) on the backbone angles. The effect of BGS is a small local defor-
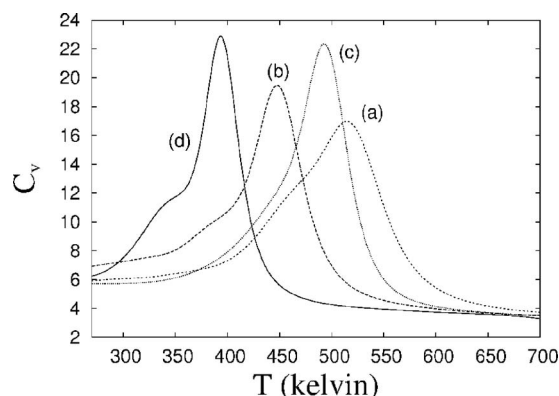
FIG. 2. Specific heat curves for (a) FSD-EY with unmodified ECEPP/3 force field, (b) FSD-EY with ECEPP/3 and only the dihedral angle corrections, (c) FSD-EY with ECEPP and only the entropic correction term, and (d) FSD-EY with ECEPP and the full correction term as in Eq. (3). The curves were obtained using the multihistogram reweighting technique of Ferrenberg and Swendsen (Ref. 22).

mation of a segment along the chain, while the downstream part of the chain from the segment also gets a small rigid-body move. The upstream part of the chain remains fixed. As implemented in SMMP, BGS is the concerted rotation of the Ramachandran angles of four consecutive residues.

III. RESULTS AND DISCUSSION

We examine four variants of the force field: plain ECEPP, ECEPP with only dihedral angle corretions, ECEPP with only entropic corrections, and ECEPP with the full corrections of Eq. (3). For each version of the force field, we performed parallel tempering simulations with 32 temperatures (32 replicas), consisting of 500 000 Monte Carlo sweeps per replica, where a sweep means as many elementary Monte Carlo updates as the number of degrees of freedom of the system. The temperatures ranged from 270 to 1000 K, initially distributed in a simple geometric series. After a preliminary run of 150 000 sweeps, the specific heat of the system was evaluated as a function of the temperature. A new temperature distribution was then estimated so that the new points divided the temperature range into regions of equal area under the approximate specific heat curve. The list of temperatures for the different force field settings can be found in the Appendix.

Figure 2 displays the specific heat curves for FSD-EY calculated from our simulations. For all four variants of our force field, we observe a peak in the specific heat that separates a high temperature phase from a low temperature phase. This peak is associated with helix formation. As shown in Fig. 3, the helix content rises sharply at a temperature corresponding to the location of the specific heat peak. Note that the correction terms shift the peak from $T_{ECEPP}=520$ K toward lower temperatures: $T_{RC}=450$ K, $T_{ent}=490$ K, and $T_{full}=400$ K. For all variations of the force field, the helical residues in the experimental structure also form a helix in the simulations. In the case of the ECEPP/3 force field without correction terms, this helix, extends to twice the length of the native helix, inhibiting the formation of the nativelike topology.
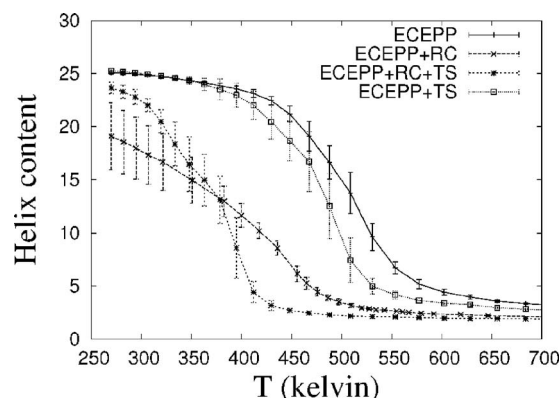

FIG. 3. Helix content vs temperature for FSD-EY for (a) unmodified ECEPP/3 force field, (b) ECEPP/3 and only the dihedral angle corrections, (c) FSD-EY with ECEPP and only the entropic correction term, and (d) ECEPP/3 and the full correction term as in Eq. (3).

This is different when the correction terms are included. The reduction in helical content in Fig. 3 is not uniform across the ensemble of structures. Rather, because of the increase of energy of the helices, a new group of structures is formed that has similar helix content as in the native state. Figure 4 displays the fraction of native configurations as a function of temperature. Here, a native configuration is defined as one with a root mean square deviation (RMSD) of less than 3.5 Å. For plain ECEPP force field and the version with only the entropic contribution included, this fraction remains below 0.1% for all temperatures and is not shown. On the other hand, when the dihedral angle correction term is included, the fraction of native configurations increases rapidly below the temperatures that mark the peak in the corresponding specific heat curve. It approaches a maximal value at $T=300$ K for the dihedral correction term and $T=350$ K for the full correction term before decreasing again. This indicates that the native state is not the global minimum in potential energy.

Note that the full correction term leads to a smaller percentage of nativelike configurations than the simple dihedral term alone. At high temperatures, the helix state is in competition with unfolded coil states which have higher entropy. This leads for simulations with only the entropic term to an onset of helix formation at a lower temperature than in the
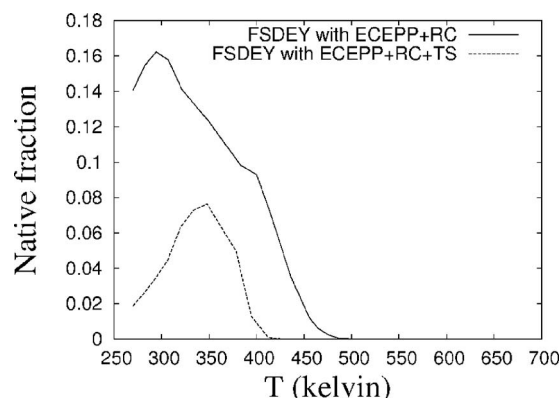

FIG. 4. Native population as a function of temperature for (a) FSD-EY under ECEPP/3 and only the dihedral angle corrections and (b) FSD-EY under the full correction term as in Eq. (3).

FIG. 5. One sequence of events which characterizes the folding process as observed in the simulations. The *C*-terminal helix forms early and often consists of the entire sequence, as that state is the global minimum of energy for our potential. When the helix partially unfolds, however, some of the unfolded parts form hydrophobic contacts with the intact part of the helix. The system then gradually rearranges into a second minimum, corresponding to the experimentally observed native state.
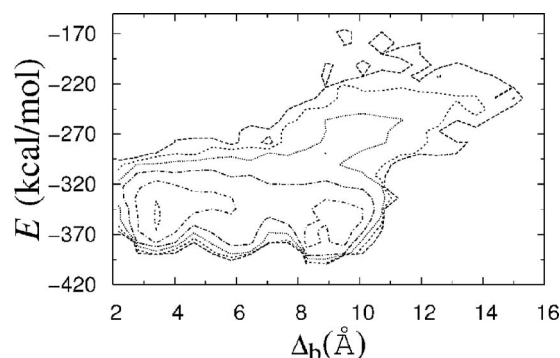

FIG. 6. Free energy landscape for FSD-EY under the variant of the force field with maximum native population at 300 K, as a function of total energy E and backbone RMSD Δ_b. The contour lines are separated by $1k_BT$. Observe that the free energy barriers are low, so that even if the native structure is a strong minimum, it has a low population. A wide variety of structures throughout this 2D projection are available at this temperature.

pure ECEPP simulations. At very low temperatures, the side chains are exposed most in the helix which therefore is favored by the entropic term. On the other hand, the dihedral angle corrections uniformly suppress helices. The effect of the full correction term is a combination of these effects.

We conclude that the simple dihedral term alone is a better correction than its combination with the entropic term. In the following, we focus on the simulations with solely the dihedral correction term. In this case, about 16% of the configurations fall at $T=300$ K within 3.5 Å of the experimentally determined structure. However, the lowest energy state is again a single long helix, as shown in the first snapshot in Fig. 5. Its total energy, $E_{tot}=-438$ kcal/mol, is about 16 kcal/mol lower than that of the lowest energy nativelike configuration found in the simulations ($E_{tot}=-422$ kcal/mol). On the average, the energy difference at $T=300$ K, between nativelike states and states with one long helix, is $\Delta E=E_{nat}-E_{helix}=20$ kcal/mol. However, after minimization the relation between the two structures is exchanged. Now the nativelike configuration with $E_{tot}=-463$ kcal/mol is energetically about as favorable as the elongated long helix with $E_{tot}=-462$ kcal/mol. We conjecture that the correction terms do indeed lead to a global minimum of energy populated by nativelike configurations but that this minimum is poorly populated and not visited in the simulation. Most nativelike configurations have higher energy than the elongated helices of Fig. 5. Both types of configurations are significantly lower in energy than the unfolded random coil states, and both form at temperatures below the specific heat peak. However, as the temperature is decreased further, the system is biased more and more in favor of the long helix which on average has lower energy. This explains the nonmonotonic behavior of the native fraction as a function of temperature, as seen in Fig. 4. Only at very low temperatures (not studied by us) we would expect a reversal of this situation.

The free energy landscape at $T=300$ K is shown in Fig. 6 as a function of backbone RMSD and energy. The nativelike structure forms a strong minimum. Another one is observed at around a RMSD of about 9 Å and separated by a free energy barrier of about 1 kcal/mol. Note that because of this projection of a high dimensional space on two dimensions the dominance of the second minimum is misleading as it contains many configurations that differ strongly between themselves. Figure 1 shows an example corresponding to the minimum associated with essentially correct structures. Note that we observe in the simulations large fluctuations in the β-hairpin part. The fluctuations are not limited to the turn

region and the dihedral angles can deviate far from the β region without unfolding the protein. This leads to the very broad nativelike minimum with configurations where the hairpin is disordered to different degrees. On the other hand, the helix remains folded all over the nativelike minimum in Fig. 6. Note that this helix includes 10–12 residues, as opposed to 9 such residues in the experimental structure. However, this difference is an artifact of our definition of helicity. The three extra helical residues fall in the experimental structure just outside the dihedral angle regions we have defined as helical, while in the simulation they fall just within these ranges.

The observed folding trajectories are consistent with the above energy landscape. Figure 5 displays snapshots of one such trajectory. A straight helical state forms along with chain collapse. One end of the helix turns out to be more prone to folding-unfolding fluctuations. If the released coil-like segment makes hydrophobic contact with the helix, it prevents immediate reformation of the long helical state. The coil-like part in hydrophobic contact with the helix rearranges itself later into a nativelike hairpin. The hydrophobic residues of the helix provide a template for the formation of this hairpin which alone is not stable. When these contacts are broken, the hairpin often unfolds, and the residues join the helix to create a long helical structure.

Our results are consistent with a recent study[19] of the folding properties of the molecule FSD-1 (PDB ID: 1FSD) using the AMBER FF03 force field, explicit water molecules, and replica exchange molecular dynamics. The molecule FSD-EY studied by us is a more stable double mutant of FSD-1. While the simulation methods and the force field used in Ref. 19 are different from those used here, the *C*-terminal helix is found to be more stable in both simulations. Only a small probability is found for the formation of the *N*-terminal β hairpin without the helix, whereas in our simulations we never observe the hairpin without the protection of the hydrophobic contacts with the helix. An earlier simulation study of thermal unfolding of both FSD-1 and FSD-EY (Ref. 20) also indicated larger conformational plasticity of the *N*-terminal β hairpin.

IV. CONCLUSIONS

Simulating with high statistics the 28-residue protein FSD-EY, we have tested correction terms to the ECEPP force field. Our data indicate that a simple dihedral term compensates best the inherent bias in ECEPP toward helical structures. With this correction term, $\approx 16\%$ of the configurations have the correct $\beta\beta\alpha$ structural motif at $T=300$ K. Analyzing folding trajectories of the protein, we find that the C-terminal helix of FSD-EY is much more stable than the N-terminal β hairpin and forms first. The hydrophobic groups of the helix provide a template that promotes the formation of the β hairpin which is never formed alone.

APPENDIX: LIST OF TEMPERATURES FOR PARALLEL TEMPERING

As described briefly in the text, the list of temperatures for parallel tempering was obtained by using short simulations for each version of the force field considered here. Initially, a simple geometric series is used for the temperatures in the range of 270–1000 K. At temperatures near the specific heat peak, the histogram of energies is widest, as the system explores both the compact states of low energies as well as some extended states of relatively high energy. Increasing the concentration of temperature points around the specific heat peak therefore facilitates more frequent transitions between the compact and extended states. To achieve this, we evaluated the specific heat curve from the preliminary runs and calculated a new temperature distribution by spacing the temperatures so that there is an equal area under the specific heat curve between successive temperature points. For the cases of pure ECEPP force field and ECEPP with only the entropic part of the correction terms in Eq. (3), no new states were observed compared to the runs with the preliminary temperature distribution. The results presented for those two cases are therefore based on runs with the following list of temperatures in kelvin: 270, 282, 294, 306, 320, 333, 348, 363, 379, 395, 412, 430, 448, 468, 488, 509, 531, 554, 577, 602, 628, 655, 684, 713, 744, 776, 810, 845, 881, 919, 959, and 1000. For the runs with ECEPP and only the dihedral angle corrections, the list of temperatures includes 270, 283, 295,307, 322, 351, 383, 400, 418, 436, 456, 465, 476, 487, 497, 510, 520, 528, 535, 543, 555, 561, 567, 579, 592, 619, 646, 705, 770, 840, 916, and 1000. For the runs with the full correction terms, the list of temperatures includes 270, 283, 296, 308, 320, 329, 338, 346, 353, 361, 368, 376, 383, 389, 396, 400, 405, 410, 416, 423, 430, 444, 464, 491, 524, 561, 603, 651, 704, 764, 832, and 910.

[1] C. A. Sarisky and S. L. Mayo, J. Mol. Biol. **307**, 1411 (2001).
[2] R. A. Abagyan and M. Totrov, J. Comput. Phys. **151**, 402 (1999).
[3] F. Eisenmenger, U. H. E. Hansmann, S. Hayryan, and C.-K. Hu, Comput. Phys. Commun. **138**, 192 (2001).
[4] F. A. Momany, R. F. McGuire, A. W. Burgess, and H. A. Scheraga, J. Phys. Chem. **79**, 2361 (1975).
[5] G. Nemethy, M. S. Pottle, and H. A. Scheraga, J. Phys. Chem. **87**, 1883 (1983).
[6] G. Nemethy, K. D. Gibson, K. A. Palmer, C. N. Yoon, G. Paterlini, A. Zagari, S. Rumsey, and H. A. Scheraga, J. Phys. Chem. **96**, 6472 (1992).
[7] M. J. Sipple, G. Nemethy, and H. A. Scheraga, J. Phys. Chem. **88**, 6231 (1984).
[8] T. Ooi, M. Oobatake, G. Nemethy, and H. A. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **84**, 3086 (1987).
[9] S. Trebst, M. Troyer, and U. H. E. Hansmann, J. Chem. Phys. **124**, 174903 (2006).
[10] A. Irbäck and S. Mohanty, Biophys. J. **88**, 1560 (2005).
[11] S. Mohanty and U. H. E. Hansmann, Biophys. J. **92**, 3573 (2006).
[12] A. Verma, S. M. Gopal, J. S. OH, K. H. Lee, and W. Wenzel, J. Comput. Chem. (to be published).
[13] T. Herges and W. Wenzel, Biophys. J. **87**, 3100 (2004).
[14] J. W. Ponder and D. A. Case, Adv. Protein Chem. **66**, 27 (2003).
[15] C. J. Geyer and E. A. Thompson, J. Am. Stat. Assoc. **90**, 909 (1995).
[16] K. Hukushima and K. Nemoto, J. Phys. Soc. Jpn. **65**, 1604 (1996).
[17] U. H. E. Hansmann, Chem. Phys. Lett. **281**, 140 (1997).
[18] G. Favrin, A. Irbäck, and F. Sjunnesson, J. Comput. Chem. **114**, 8154 (2001).
[19] W. Li, J. Zhang, and W. Wang, Proteins **67**, 338 (2007).
[20] H. Lei and Y. Duan, J. Chem. Phys. **121**, 12104 (2004).
[21] W. L. DeLano, The PYMOL molecular graphics system, DeLano Scientific, San Carlos, CA, 2002.
[22] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **63**, 1195 (1989).