J|A|C|S
ARTICLES
Published on Web 01/21/2004

Polypeptide Folding Using Monte Carlo Sampling, Concerted Rotation, and Continuum Solvation

Jakob P. Ulmschneider and William L. Jorgensen*

Contribution from the Department of Chemistry, Yale University, New Haven, Connecticut 06520-8107

Received August 12, 2003; E-mail: william.jorgensen@yale.edu

Abstract: An efficient concerted rotation algorithm for use in Monte Carlo statistical mechanics simulations is applied to fold three polypeptides, U(1−17)T9D, α_1, and trpzip2, which exhibit native β-hairpin and α-helix folds. The method includes flexible bond and dihedral angles, and a Gaussian bias is applied with driver bond and dihedral angles to optimize the sampling efficiency. Solvation in water is implemented with the generalized Born (GBSA) model. The computed lowest-energy manifolds for the folded structures of the two β-hairpins agree closely with the corresponding NMR structures. In the case of the α_1 peptide, the folded α-helical state, which is observed as oligomers in concentrated solution and crystals, is not stable in isolation. The computed preference for random coil structures is in agreement with NMR experiments at low concentration. The fact that native states can be located on high dimensional energy surfaces starting from extended conformations shows that the present methodology samples all relevant parts of the conformational space. The OPLS-AA force field with the GBSA solvent model was also found to perform well in leading to clear energetic separation of the correctly folded structures from misfolded structures for the two peptides that form β-turns.

Introduction

The folding of proteins into their native structure is one of the most challenging and interesting problems of molecular biology. In addition to many experimental efforts, computer simulations have been performed at various levels of complexity. Monte Carlo (MC) lattice simulations, where the protein is approximated as a series of beads on a lattice, have been used for a long time due to their low computational costs (for some recent examples, see refs 1−4). More realism is obtained with minimalist off-lattice models, where the side chains are reduced to spheres or where united atom representations are used.[5−7] The simplicity and computational efficiency of these models have made it possible to simulate thousands of folding−unfolding events to obtain a detailed statistical description of the folding process. At the highest level of detail and computational cost are all-atom representation models with implicit or explicit representation of the solvent. To study a complete folding pathway using traditional molecular dynamics (MD) or MC has, until recently, been beyond the computational possibilities for any but the smallest systems. However, a wealth of other methods to elucidate folding mechanisms have been employed including minimization techniques,[8] configurational space annealing,[5,9] simulated annealing,[10] multicanonical simulations,[11,12] elevated temperature MD unfolding studies,[13−15] and more recently, replica exchange methods[16,17] and ensemble dynamics.[18,19]

Due to the high computational cost of explicit solvent representation, there has been increased interest in the use of implicit solvation models, which reduce the computational burden through a continuum treatment of the solvent. Of these, the generalized Born (GBSA) solvent model has been widely applied because it is computationally efficient and physically understandable. Originally developed by Still et al.,[20] the model is an extension of the Born treatment of ionic solvation to solutes containing any set of charged sites and having arbitrary molecular shapes. The GBSA treatment of electrostatics includes both the screening of charge−charge interactions and the "self" energies related to burial of charges or dipoles in the low-

(1) Mirny, L.; Shakhnovich, E. *Ann. Rev. Biophys. Biomol. Struct.* **2001**, *30*, 361−396.
(2) Dinner, A. R.; Sali, A.; Smith, L. J.; Dobson, C. M.; Karplus, M. *Trends Biochem. Sci.* **2000**, *25*, 331−339.
(3) Zhdanov, V. P. *Europhys. Lett.* **1998**, *42*, 577−581.
(4) Dimitrievski, K.; Kasemo, B.; Zhdanov, V. P. *J. Chem. Phys.* **2000**, *113*, 883−890.
(5) Lee, J.; Liwo, A.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 2025−2030.
(6) Derreumaux, P. *Phys. Rev. Lett.* **2000**, *85*, 206−209.
(7) Klimov, D. K.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 2544−2549.
(8) Wales, D. J.; Scheraga, H. A. *Science* **1999**, *285*, 1368−1372.
(9) Lee, J. Y.; Scheraga, H. A. *Int. J. Quantum Chem.* **1999**, *75*, 255−265.
(10) Liu, Y. X.; Beveridge, D. L. *Proteins* **2002**, *46*, 128−146.
(11) Alves, N. A.; Hansmann, U. H. E. *J. Chem. Phys.* **2002**, *117*, 2337−2343.
(12) Dinner, A. R.; Lazaridis, T.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9068−9073.
(13) Daggett, V.; Fersht, A. *Nature Rev. Mol. Cell Biol.* **2003**, *4*, 497−502.
(14) Tirado-Rives, J.; Jorgensen, W. L. *Biochemistry* **1993**, *32*, 4175−4186.
(15) Lee, J.; Shin, S. M. *Biophys. J.* **2001**, *81*, 2507−2516.
(16) Zhou, R. H.; Berne, B. J. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12777−12782.
(17) Zhou, R. H.; Berne, B. J.; Germain, R. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 14931−14936.
(18) Zagrovic, B.; Sorin, E. J.; Pande, V. *J. Mol. Biol.* **2001**, *313*, 151−169.
(19) Pande, V. S.; Baker, I.; Chapman, J.; Elmer, S. P.; Khaliq, S.; Larson, S. M.; Rhee, Y. M.; Shirts, M. R.; Snow, C. D.; Sorin, E. J.; Zagrovic, B. *Biopolymers* **2003**, *68*, 91−109.
(20) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.

dielectric interior of macromolecules. The model also features a sum over atomic solvent-accessible surface areas (SA) to estimate the free energy of cavity formation in the solvent and solute−solvent van der Waals' interactions. Continuum representation of the solvent can, of course, be criticized since the solvent is modeled as an isotropic linear dielectric, with a time- and space-averaged bulk value of the dielectric constant. Using explicit solvation, local regions show fluctuations in the effective dielectric constant. Also, bound surface water around a protein is not expected to have the same dielectric constant as bulk water. In addition, continuum models fail to predict temperature and pressure effects correctly and omit frictional effects of water on solute dynamics. The GBSA method further suffers from the neglect of field boundary conditions at the dielectric interface (Coulomb−Field approximation). Despite all these concerns, results from the GBSA model have consistently been shown to compare favorably with experimental data and numerical solution of Poisson's equation, and it reproduces accurately relative free energies of different peptide conformations.[21] GBSA is certainly superior to earlier, simpler alternatives using just surface-area terms or a distance-dependent dielectric model,[21] and its further exploration in protein folding studies is of interest.

Any energy function used for folding studies has to be able to identify the native fold from the many alternative non-native folds. Indeed, the question of whether standard molecular mechanics force fields in conjunction with generalized Born solvation models can be used for successful structure scoring has recently been addressed. In an extensive comparison with large decoy sets, the OPLS-AA[22] force field combined with the surface-generalized Born model[23] was shown to identify correctly the native state of several large proteins.[24] While many recent efforts have focused on determining the free energy surface[12,16,25] or finding the global minimum with noncanonical methods, some recent studies have used MD simulations coupled with the GBSA model to fold small polypeptides.[26−28] The success of this approach is impressive and led us to consider an alternative ab initio method that combines MC statistical mechanics, our newly developed improved concerted rotation sampling procedure, and the GBSA model.[29] The combination of MC with GBSA has potential advantages over MD. In particular, energy derivatives are not needed, including the costly ones for the GBSA free energy. Furthermore, MC moves can take better advantage of the implicit nature of the solvent by enabling large conformational changes to cross efficiently over energy barriers. On the other hand, the use of MC to simulate polymer dynamics requires nontrivial move sets. This is due to the fact that simple MC variation of individual degrees of freedom in the backbone leads to global conformational changes accompanied by intramolecular steric clashes for any but the smallest moves. The problem was first addressed by Dodd et al.[30] for polymers and Knapp and Irgensdefregger[31,32] for proteins by introducing local backbone moves termed *concerted rotations* that change a number of consecutive backbone torsion angles but leave the rest of the main chain unchanged.

Concerted rotations successfully avoid global conformational changes and have the additional advantage that only a small part of the time-consuming nonbonded energy terms have to be updated for every MC move. However, they introduce new problems. Due to the constraints of the methods, only the torsion angles of the backbone were taken as variable degrees of freedom. The lack of backbone bond stretching and angle bending leads to artificial rigidity and reduced conformational transitions,[33] similar to the situation with MD simulations applying holonomic constraints to bond lengths and angles. Small variations of bond lengths and, especially, bond angles reduce considerably rotational barriers.[34] The use of a standard molecular mechanics force field is, therefore, problematic, and modified torsion potentials have been proposed as a solution.[34] MC simulations were also performed with these concerted rotation algorithms to fold polypeptides in helix-turn-helix[35] and β-hairpin[36] motifs. To improve directly the flexibility of the sampling, we reported an improved concerted rotation scheme, named concerted rotation with bond angles (CRA).[29] The scheme has both the torsion angles and bond angles of the backbone flexible and uses a Gaussian bias to incorporate the increased number of degrees of freedom. The biasing and higher segment closure probability make the CRA method faster and more efficient than the prior alternatives.[29] In this study, we demonstrate the strength of the combined MC/CRA/GBSA approach for folding polypeptides and identifying their native structures in aqueous solution. Preliminary studies on polyalanine chains folding into α-helices were encouraging.[29] Results for three diverse polypeptides, which are known to favor β-hairpin and α-helical structures, are now reported.

Computational Procedure

The simulations were run with the MCPRO program[37] modified to include the concerted rotations, as detailed in the original report.[29] Each simulated system consisted of just a single copy of the polypeptide. Normal protonation states were adopted for pH 7, i.e., deprotonated carboxylic acids and protonated amines and guanidines; there were no histidines in the structures. The potential energy was evaluated with the OPLS-AA force field,[22] and the MC simulations used Metropolis sampling mostly at temperatures of 30−50 °C, as detailed below. The full potential energy was evaluated with no cutoffs for the nonbonded interactions and with a dielectric constant of 1 for the Coulombic interactions. The utilized GBSA method was the analytical model developed by Qiu, Still, and co-workers.[38] The electrostatic energy and,

(21) Bashford, D.; Case, D. A. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129−152.
(22) Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.
(23) Ghosh, A.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B* **1998**, *102*, 10983−10990.
(24) Felts, A. K.; Gallicchio, E.; Wallqvist, A.; Levy, R. M. *Proteins* **2002**, *48*, 404−422.
(25) Bursulaya, B. D.; Brooks, C. L. *J. Phys. Chem. B* **2000**, *104*, 12378−12383.
(26) Simmerling, C.; Strockbine, B.; Roitberg, A. E. *J. Am. Chem. Soc.* **2002**, *124*, 11258−11259.
(27) Jang, S.; Shin, S.; Pak, Y. *J. Am. Chem. Soc.* **2002**, *124*, 4976−4977.
(28) Chowdhury, S.; Zhang, W.; Wu, C.; Xiong, G. M.; Duan, Y. *Biopolymers* **2003**, *68*, 63−75.
(29) Ulmschneider, J. P.; Jorgensen, W. L. *J. Chem. Phys.* **2003**, *118*, 4261−4271.
(30) Dodd, L. R.; Boone, T. D.; Theodorou, D. N. *Mol. Phys.* **1993**, *78*, 961−996.
(31) Knapp, E. W. *J. Comput. Chem.* **1992**, *13*, 793−798.
(32) Knapp, E. W.; Irgensdefregger, A. *J. Comput. Chem.* **1993**, *14*, 19−29.
(33) Bruccoleri, R. E.; Karplus, M. *Macromolecules* **1985**, *18*, 2767−2773.
(34) Sartori, F.; Melchers, B.; Bottcher, H.; Knapp, E. W. *J. Chem. Phys.* **1998**, *108*, 8264−8276.
(35) Hoffmann, D.; Knapp, E. W. *J. Phys. Chem. B* **1997**, *101*, 6734−6740.
(36) Rabenstein, B.; Hoffmann, D.; Knapp, E. W. *AIP Proceedings of the Third International Symposium on Biological Physics*; Third International Symposium on Biological Physics 1998; Santa Fe, NM, September 20−24, 1998; American Institute of Physics: New York, 1999; pp 54−68.
(37) Jorgensen, W. L.; Tirado-Rives, J. *MCPRO*, v 1.68; Yale University: New Haven, CT, 2002.
(38) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. *J. Phys. Chem. A* **1997**, *101*, 3005−3014.

therefore, the Born radii are recomputed for every MC configuration; the constituent atomic radii are taken from the OPLS-AA force field ($r = 0.5\sigma$) except in the case of hydrogens for which radii of 1.15 Å are assigned, as in the original study.[38] The nonpolar contribution to the solvation free energy was calculated as in the original method by Still et al.[20] to be proportional to the total solvent-accessible surface area (SASA) with an effective surface tension of 7.3 cal/mol-Å^2. The SASA was computed using a probe radius of 1.4 Å. For the MC simulations of polypeptides, given that the SASA calculation is time-consuming, that SASA is slowly varying, and that the contributions of the SASA term to the free energies are relatively small, the SASA was updated every 30 MC steps. This approximation can also be found in other studies.[39] Attempted backbone moves were made every fourth configuration; the remainder was single side-chain moves, which are rapid. Although the lengths for the different MC simulations are provided below, typical runs of 100×10^6 (100 M) configurations for the 17-residue U(1−17)T9D peptide required 5 days on a 2.4 GHz Pentium IV.

It was verified that our GBSA implementation yields nearly identical free energies of hydration as in the original study;[38] for the original 35 organic molecules, both implementations yield correlation coefficients r^2 of 0.9 and mean unsigned errors (mue) of 0.9−1.0 kcal/mol vs experiment. Small discrepancies arise from the use of slightly different atomic radii and atomic changes.[20] We then extended the test set to over 400 organic compounds; comparison with experimental data did not reveal significant flaws especially in the treatment of nonpolar solutes for which the SA term is dominant. For polar solutes, the contribution of this term to the total free energy of hydration tends to be small, and the electrostatic term is dominant. The mue does rise to 1.2 kcal/mol owing entirely to poor performance for the 14 nitro compounds in the dataset; the mue is 0.9 kcal/mol without them. Our implementation of GBSA was also compared to another common GBSA method, the pairwise model of Hawkins, Cramer, and Truhlar[40] as implemented in the AMBER 7 program suite.[41] In this case, the free energies of hydration were calculated with both methods for 50 evenly spaced snapshots from one of our MC runs for the U(1−17)T9D peptide; the two sets of results correlate with an r^2 of 0.99. This is notable, given the differences in the methods and parametrization.

The principal experimental data for comparison with the present simulation results are the structures of the polypeptides as obtained from detailed NMR studies. All three polypeptides have been examined by NMR in aqueous solution, and structures for two of them have been deposited in the Protein Data Bank (PDB), as elaborated on below.

Results

***β*-Hairpin U(1−17)T9D.** The first system studied was a 17-residue polypeptide, U(1−17)T9D,[42] derived from the small globular protein ubiquitin, with the sequence MQIFVKTLDGK-TITLEV. A NMR study has revealed that it forms a β-hairpin in aqueous solution (Protein Data Bank (PDB) code 1e0q).[42] This system has recently been examined with MD simulations using the GBSA solvation model and the CHARMM[43] force field at elevated temperature, 360 K.[27] In our case, a series of eight MC runs was performed at 30 °C (303 K), starting from completely extended conformations. Figure 1 shows the development of major observables over the course of run 1, the total energy (intramolecular energy plus solvation free energy), the number of backbone interstrand hydrogen bonds, the radius of gyration, and the backbone (C, C_α, N) root-mean-square deviation (RMSD) to the NMR structure (conformer 8 of the 27 structures deposited[42]).

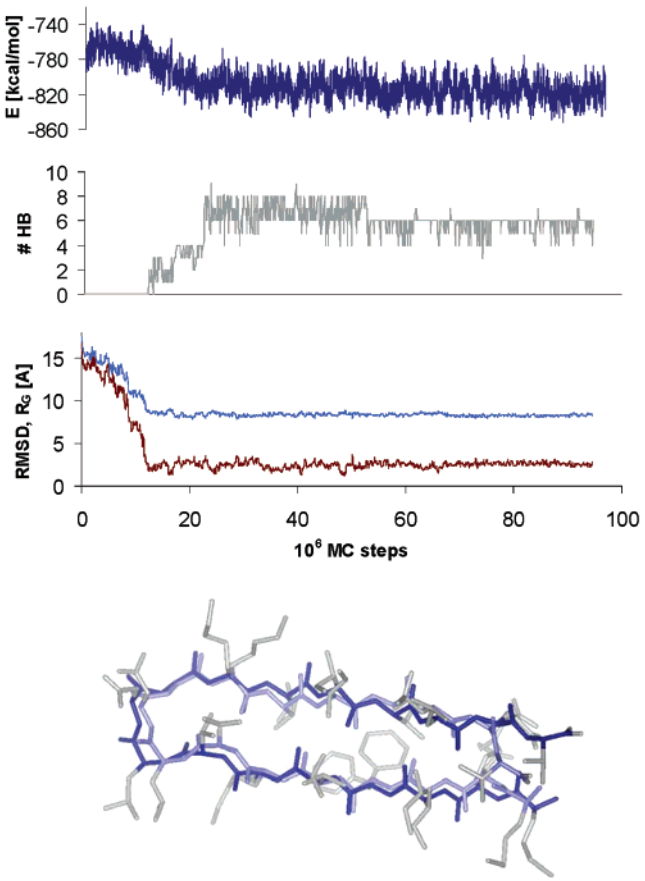

Figure 1. Total energy, number of backbone hydrogen bonds, radius of gyration, and RMSD to the NMR structure as a function of MC steps for run 1 of the eight simulations of U(1−17)T9D. An overlay of a representative structure from the folded phase with the NMR structure is shown.

All runs led to stable β-hairpin conformations within 20−80 M configurations. In all cases, a rapid relaxation of the extended structure to a more compact state characterized by a twisted backbone is observed. This phase exhibits high fluctuations, and its length varies from just a few million MC steps to at most ca. 70 M configurations. During this period, the structure remains extended with backbone RMSDs of 5−15 Å and a radius of gyration of 12−15 Å. The subsequent phase is characterized by a sudden transition to a U-shaped structure and formation of a turn in the middle of the peptide. The side chains of the opposing β-strands come into contact, and there is a fast rearrangement characterized by a significant drop in energy, a build-up of the backbone hydrogen bonds, and formation of packing between the side chains. Once the backbone hydrogen bonds have formed, the structure remains stable for the rest of the simulation. The backbone RMSD to the native structure ranges between 2.5 and 5 Å for the folded phase, and the system remains compact with gyration radii of 8.6−9.3 Å. There is very little subsequent variation of any observable.

The resulting averages over the stable folded phase are shown in Table 1, sorted by increasing RMSD. There are clear correlations between the backbone RMSD, the total energy of

(39) Zhu, J. A.; Shi, Y. Y.; Liu, H. Y. *J. Phys. Chem. B* **2002**, *106*, 4844−4853.
(40) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 19824−19839.
(41) *AMBER*, v7; University of California: San Francisco, CA, 2003.
(42) Zerella, R.; Chen, P. Y.; Evans, P. A.; Raine, A.; Williams, D. H. *Protein Sci.* **2000**, *9*, 2142−2150.
(43) MacKerell, A. D.; Banavali, N.; Foloppe, N. *Biopolymers* **2000**, *56*, 257−265.

Table 1. Averages over the Folded Phase during the Eight Simulations of U(1−17)T9D for Total Energy, RMSD to the NMR Structure, Number of Interstrand Backbone Hydrogen Bonds, and Radius of Gyration

run	⟨*E*⟩ [kcal/mol]	⟨RMSD⟩ [Å]	⟨# HB⟩	⟨R_G⟩ [Å]
1	−815.0 11.5	2.5 ± 0.4	6.1	8.4 ± 0.2
2	−799.0 11.1	2.7 ± 0.2	5.1	8.6 ± 0.2
3	−801.4 11.2	2.8 ± 0.3	5.1	8.6 ± 0.2
4	−796.5 10.6	3.9 ± 0.3	3.4	8.6 ± 0.2
5	−799.7 10.6	4.2 ± 0.2	3.0	8.5 ± 0.2
6	−799.4 9.2	5.0 ± 0.3	1.9	9.3 ± 0.2
7	−792.7 11.2	5.1 ± 0.6	2.0	9.3 ± 0.2
8	−794.4 10.8	5.1 ± 0.2	2.0	9.2 ± 0.2

Table 2. Lengths of Stable Backbone Hydrogen Bonds for the U(1−17)T9D Polypeptide[a]

	NMR data [Å]	MC run 1 [Å]
Met 1 (O)−Val 17 (N)	4.0 ± 0.4	3.1 ± 0.3
Leu 15 (O)−Ile 3 (N)	3.4 ± 0.2	3.3 ± 0.3
Ile 3 (O)−Leu 15 (N)	3.13 ± 0.08	3.2 ± 0.3
Ile 13 (O)−Val 5 (N)	3.7 ± 0.4	3.1 ± 0.2
Val 5 (O)−Ile13 (N)	3.8 ± 0.3	3.2 ± 0.5
Lys 11 (O)−Thr 7 (N)	3.6 ± 0.2	3.2 ± 0.8

[a] All bonding partners observed in the NMR study were found in MC simulation run 1. The averages were taken over the completely folded part of the trajectory.

the system, and the amount of backbone hydrogen bonding. The structure with by far the lowest energy is also structurally closest to the NMR findings with formation of the correct backbone hydrogen bonds and the correct turn structure. The seven other runs also find β-hairpins with a slightly different structure and are sampling higher energy local minima. Looking at the trajectories, it is clear that establishment of a turn in the middle of the chain is the key event leading to the formation of the hairpin. All eight simulations established a hairpin with a turn at the chain center, and no simulations yielded a turn at any other position or any helical structures. Of considerable interest is the role of Gly10, whose flexibility promotes establishment of the turn region in its vicinity. The NMR conformation shows a type I β-turn with the Gly at the fourth position in the turn sequence Thr-Leu-Asp-Gly, but there could be competing alternatives, notably due to the preference of Gly to be at position three in a β-turn.[44] This alternative structure would have the opposing chains displaced by one residue and was suggested by Zerella et al.[42] but was not observed in the NMR study.

There is remarkable similarity between the stable structure obtained in the simulation converging to the lowest energy (run 1) and the structure from the NMR measurements,[42] as illustrated by their superposition in Figure 1. A stable type I β-turn was obtained in the simulation for the Thr-Leu-Asp-Gly segment in perfect agreement with the NMR findings. All six backbone hydrogen bonds found in the NMR structure are highly populated, namely Met1(O)−Val17(N), Lys11(O)−Thr7(N) and the pairs of hydrogen bonds for Ile3−Leu15 and Val5−Ile13. The NMR measurements indicated slightly longer hydrogen bonds due to the more dynamic behavior of the β-hairpin in solution as compared to that of the corresponding β-hairpin in the crystal structure of ubiquitin. The backbone hydrogen bond lengths averaged over the folded phase of run 1 are shown in Table 2 and compared to the averages from the NMR study; the structures are clearly the same with somewhat shorter contacts in the simulation. The side-chain conformations in the NMR measurements were well defined for some residues, all showing local packing interactions. For other residues the structure was less defined, indicating considerable flexibility of the peptide in solution. The corresponding behavior of the well-defined side chains during the stable phase of run 1 reproduces these findings. A small difference is looser packing of Ile3 and Val5 compared to that in the NMR structures. The simulations show the methyl group of Thr12 making contact with the side-chain methylene groups of Lys6, and the methyl group of Thr14 is well packed against the ring of Phe4, in agreement with experiment.[42] On very few occasions during the simulation, the Thr14 side chain showed an alternative structure by pointing the β-methine group toward the ring of Phe4, an observation also noted during the NMR refinement.[42] An interesting question concerns the non-native residue Asp9, which is thought to be responsible for the increased stability of the T9D peptide. The NMR study found an NOE between the side chains of Asp9 and Lys11 as well as pH dependence of the hairpin stability. This suggests that Asp9 needs to be deprotonated and that salt-bridge formation between Asp9 and Lys11 is stabilizing. Indeed, during the simulation, frequent contact was observed between the NH_3^+ of Lys11 and the COO^- of Asp9.

It is also of interest to compare the present MC structure with the MD results of Jang et al.,[27] which yielded five folded β-hairpins at 360 K using a different force field. Jang et al. report a backbone RMSD of 1.36 Å for one computed structure to the average NMR structure. In our simulation, the RMSD is as low as 1.2 Å, but the average RMSD over the folded part of the trajectory is 2.5 Å. While all backbone interstrand hydrogen bonds remain intact in the converged phase of the MC run, there is still considerable twisting and fluctuation in the β structure.

Although the similarities of the lowest-energy minimum encountered in run 1 to the NMR structure are encouraging, the nature of the final structures from the seven other simulations warrants consideration as well. Their average potential energies are all −797 ± 4 kcal/mol, which is 15−20 kcal/mol above the structure from run 1. All seven runs yield stable β-hairpins, and there is close similarity between these misfolded structures. All show a type II β-turn with Gly10 at the third position in the Leu-Asp-Gly-Lys bend. Thus, the interstrand contacts are out of register with respect to the native state, and the number of backbone hydrogen bonds is lower, as can be seen in Table 1. Due to the flip of one of the strands of the hairpin, the side-chain structure is very different, and the seven runs exhibit various non-native contacts between the residues. The fluctuations in the final state indicate that the system is trapped in series of broader local minima. No complete unfolding of an initially formed β-hairpin was observed during the present simulations. Due to the large energetic advantage of the native state and no obvious entropic disadvantage, in longer simulations or at higher temperatures the misfolded peptides should leave their higher-energy local minima and ultimately find the correct native backbone structure. This needs to be studied systematically to devise optimal procedures for convergence to the native state.

The folding mechanism of small β-hairpin peptides is being investigated actively as a prototype for the larger protein folding

(44) Hutchinson, E. G.; Sessions, R. B.; Thornton, J. M.; Woolfson, D. N. *Protein Sci.* **1998**, *7*, 2287−2300.

problem. Numerous recent efforts have centered on a β-hairpin of similar size, the C-terminal hairpin from protein G (PDB code 1gb1).[12,16−18,45−47] The simulations have employed replica exchange methods (REM),[16,17] multicanonical Monte Carlo sampling,[12] and ensemble dynamics[18] in order to sample exhaustively the conformational space accessible to the peptide. From an analysis of the free energy surfaces, several folding mechanisms have been proposed. In the hydrogen-bond-centric view,[45,46] folding is initiated by formation of the turn, followed by propagation of hydrogen bonds to the chain ends with late formation of the hydrophobic side-chain clusters that stabilize the folded hairpin. In the hydrophobic-core-centric alternative,[47] the protein first collapses to an intermediate compact state characterized by few hydrogen bonds but a formed hydrophobic core. The system then folds into the native state by forming the interstrand hydrogen bonds. Several recent studies[17,18] argue that a mix of these two extreme cases is probably most realistic, given the lack of observing a fast zipping transition expected from the first model or of finding a heavily populated state with a hydrophobic core but no hairpin hydrogen bonds expected from the second model. In the blended view, the collapse and partial hydrogen bond formation occur at the same time. On this pathway, considerable time can be spent visiting partially folded structures that are compact and have some of the native hydrogen bonds.

The mixed picture of hairpin folding is consistent with the pattern that is followed in the present MC simulations, although it is cautioned that the folding pathway from a MC run reflects no time element and is affected by the choice of sampling algorithm. In all runs, there is a clear transition between an extended state characterized by high RMSD (>10 Å), high R_G, and no interstrand hydrogen bonds, and a folded compact state characterized by a RMSD of 2.5−5 Å, a R_G of 8.6−9.3 Å, and significant establishment of interstrand hydrogen bonding. The transition occurs sharply over a MC "time" scale of 2−10 M configurations without the occurrence of any detectable long-lived intermediate state. No α-helical structures were encountered in any of the eight simulations, and there was no persistent helicity observed for any individual residues, in agreement with the NMR study. There was also an absence of any compact random-coil structures. The folded ensemble consists entirely of β-structures, and the unfolded ensemble is dominated by twisted and bent extended conformations. Experimental results indicate that the folded hairpin is highly populated, with an estimate of ∼64% at room temperature given by Zerella et al.[42] The fact that all extended starting structures transition rapidly to hairpins in the MC runs is consistent with the experimental findings. Although misfolded hairpins were obtained, the results for run 1 indicate that these metastable structures are not required intermediates on the folding pathway. The misfolded structures were also easy to identify in this case by their higher average energies.

α-Helical Polypeptide α_1. The present simulation technique was next applied to the α_1 polypeptide that was designed to form α-helical bundles by Ho and DeGrado.[48] The 12-residue peptide has the sequence ELLKKLLEELKG. An amphiphilic design was followed with leucine residues on one face of the helix stabilizing hydrophobic helix−helix contacts and charged glutamic acid and lysine residues facing the solvent on the opposing side. The crystal structure has been determined by Hill et al.[49] (PDB code 1al1), revealing antiparallel dimers in tetramers and hexamers with the monomer units α-helical except for the terminal glycine. Crystal structures at higher pH find similar oligomers containing completely helical monomer units.[50] NMR studies show that at high concentrations this state also exists in solution.[51] However, it was observed that at lower concentrations, where no formation of oligomers takes place, the monomer unit seems to be a random coil. There have been computational studies on the series of helical bundle proteins designed by DeGrado et al. involving reduced protein and lattice models and confirming the assembly into oligomers.[52,53] An α_1-like monomer has also recently been part of a protein folding study employing simulated annealing with an all-atom force field and the GBSA solvation model.[10] The lowest free energy ("folded") state was shown to be the same helical state observed in the crystal structure. Using our methodology, it is of interest to investigate whether this free energy minimum is highly populated in solution in contradiction to the experimental results at low concentration.[51]

A total of four MC runs were started from completely extended conformations for lengths of 350−800 M configurations at a slightly elevated temperature of 40 °C (313 K), which was found to be more optimal in this case than runs at 30 °C for fast exploration of the conformational space. Our analysis of the results focuses on the main-chain helical content and the diversity of the structures. We define a residue to be helical when the Φ, Ψ torsion angles are within a range (±30°, ±25°) of their ideal α-helical values of $\Phi = -57°$ and $\Psi = -47°$. Main-chain hydrogen bonds are defined by an O···HN distance below 2.8 Å and a CO···H angle between 120° and 180°. The development of the helicity per residue and the number of α-helical (i, $i + 4$) main-chain hydrogen bonds over the course of one of the trajectories is shown in Figure 2. It is clear that there is substantial helical structure with partial helices forming and dissolving many times during the simulation. The position of the helical segments shows a preference for the center with the chain ends frayed. No 3_{10}-helix or π-helix formation was observed. During the four simulations, completely folded α-helical conformations were encountered several times (<0.5 Å backbone RMSD to a reference complete right-handed α-helix), and they were stable for 10−30 M MC configurations.

A cluster analysis was carried out to assess the main secondary structural motifs populated during the simulations. Since pairwise clustering becomes computationally costly for large coordinate sets, the structures were taken every millionth MC step and were superimposed using main-chain least-squares superimposition with a similarity cutoff of 1.5 Å. The pairwise method of Daura et al.[54] was employed. A total of 606 clusters were found. Since most clusters are sparsely populated and do

(45) Munoz, V.; Thompson, P. A.; Hofrichter, J.; Eaton, W. A. *Nature* **1997**, *390*, 196−199.
(46) Munoz, V.; Henry, E. R.; Hofrichter, J.; Eaton, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 5872−5879.
(47) Garcia, A. E.; Sanbonmatsu, K. Y. *Proteins* **2001**, *42*, 345−354.
(48) Ho, S. P.; DeGrado, W. F. *J. Am. Chem. Soc.* **1987**, *109*, 6751−6758.
(49) Hill, C. P.; Anderson, D. H.; Wesson, L.; DeGrado, W. F.; Eisenberg, D. *Science* **1990**, *249*, 543−546.
(50) Prive, G. G.; Anderson, D. H.; Wesson, L.; Cascio, D.; Eisenberg, D. *Protein Sci.* **1999**, *8*, 1400−1409.
(51) Ciesla, D. J.; Gilbert, D. E.; Feigon, J. *J. Am. Chem. Soc.* **1991**, *113*, 3957−3961.
(52) Smith, A. V.; Hall, C. K. *J. Mol. Biol.* **2001**, *312*, 187−202.
(53) Sikorski, A.; Kolinski, A.; Skolnick, J. *Biophys. J.* **1998**, *75*, 92−105.
(54) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. *Angew. Chem., Int. Ed.* **1999**, *38*, 236−240.

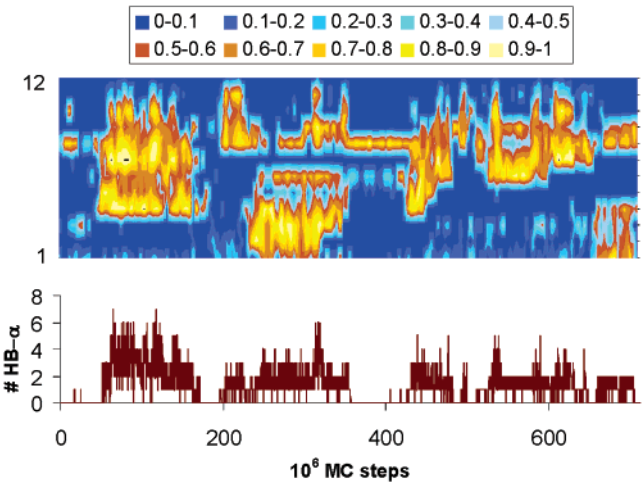

Figure 2. Helical content per residue (top) and number of helical i, $i + 4$ backbone hydrogen bonds (bottom) over the course of one of the MC trajectories for the α_1 peptide.

not represent the main features of the simulations, only the 200 most populated clusters were further analyzed. The clusters with similar secondary structural motifs were grouped together, and the results are illustrated in Figure 3 and summarized in Table 3. It is clear that the system explores most conformational possibilities without getting trapped in a local minimum for too long at this temperature, 40 °C. Structures showing helical content make up 43% of the analyzed structures (Figure 3 a−d). However, the majority of these structures exhibits only one helical turn, and the high number of clusters is due to the variation in position of the helical segment. Two-turn helices make up about 15% of the structures, and only 1% shows a completely folded helix. This indicates that the fully helical monomer of α_1 is not stable at this temperature on its own.

Table 3. Cluster Analysis of All Four MC Runs for the α_1 Polypeptide[a]

	structures (clusters)	occurrence (%)
single-turn helix	596 (68)	27
two-turn helix	332 (20)	15
full helix	20 (2)	1
β structures	458 (29)	21
random coil	343 (81)	16

[a] A total of 606 clusters was found. The first 200 most populated clusters are represented, classified according to secondary structure.

There is a significant population, 21%, of β-like structures, and 16% of the structures show no significant structural features and are classified as random coil. There are an additional 20% of the structures that fall into less populated clusters and can also mostly be categorized as random coils.

The above results indicate that the system is a random coil with some population of partially helical and β-structures. There is no dominant stable conformation at this temperature in agreement with the NMR findings.[51] Nevertheless, the exhibited helical behavior is a consequence of the amphiphilic design that works even for the monomer by favoring partial helices to bring the hydrophobic leucine residues in contact. It seems that a stable structure can only be gained by hydrophobic clustering of two or more monomers. The completely helical state and the nearly completely helical state encountered in the simulated annealing study[10] were visited several times during the simulations, but were not stable.

Tryptophan Zipper. Tryptophan zippers (TrpZip) is the name given to a series of small peptides recently synthesized by Cochran et al.[55] Despite their size of only 12−16 residues, they form remarkably stable β-hairpins in aqueous solution characterized by a structural motif of tryptophan−tryptophan cross-strand pairs. They are considered to be the smallest

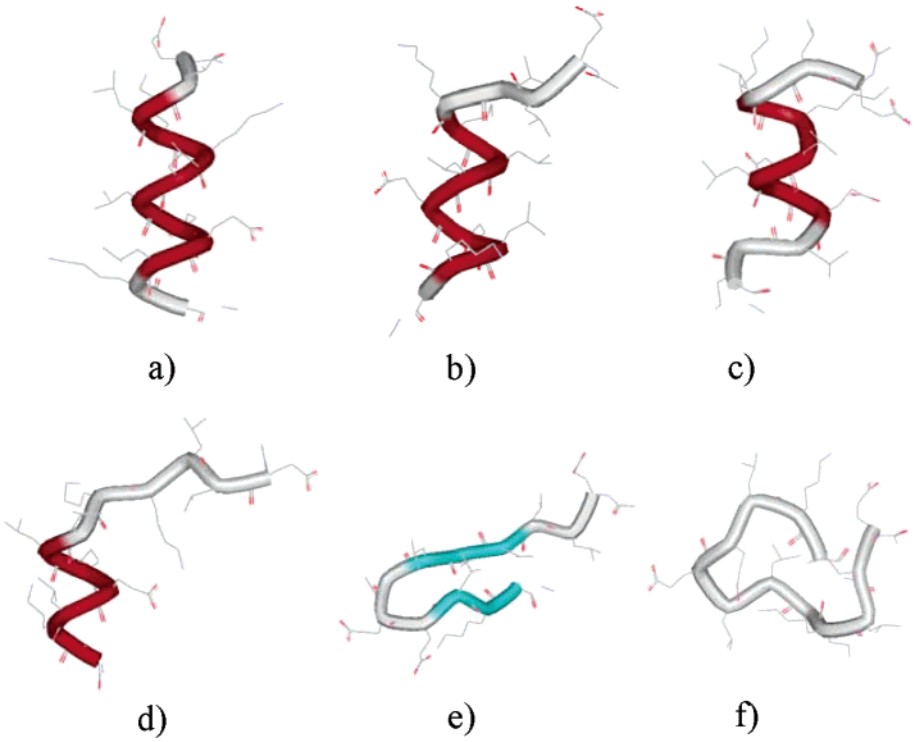

Figure 3. Representative structures from the most populated clusters for the α_1 peptide. They are classified according to secondary structure, (a) full helix, (b) two-turn helix with loose end, (c) two-turn helix with both ends frayed, (d) single-turn helix, (e) β-structures, and (f) random coils.

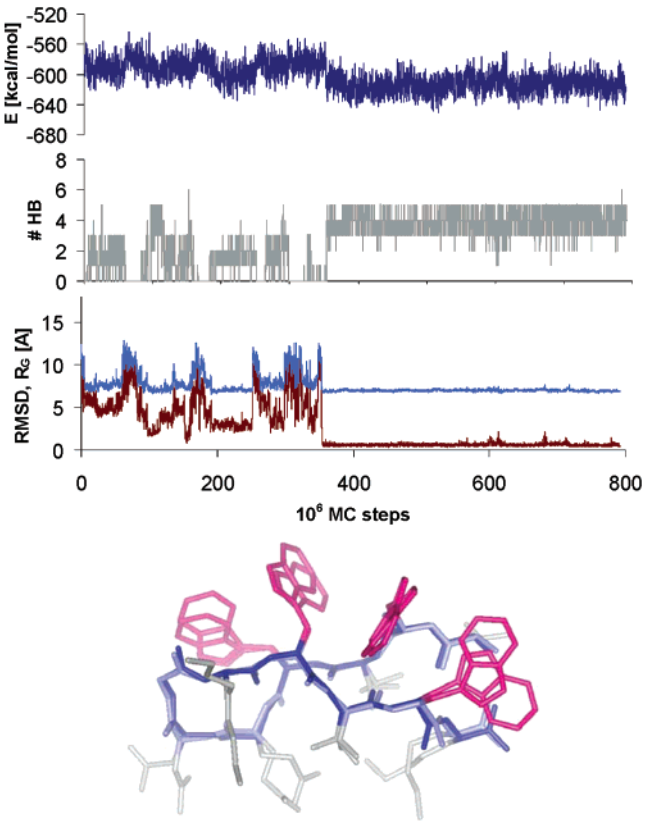

Figure 4. Total energy, number of backbone hydrogen bonds, radius of gyration, and RMSD to the NMR structure as a function of MC configurations for run 1 of the eight simulations of trpzip2. An overlay of a representative structure from the folded phase with the NMR structure is shown.

peptides exhibiting a unique tertiary β-fold and the thermodynamic properties of larger folded proteins.[55] As such they represent a good further model system, and the tryptophan zipper, trpzip2 (sequence SWTWENGKWTWK), was chosen for simulation due to its high stability. CD spectroscopy and NMR experiments reveal a β-hairpin with a type I′ β-turn at the Asn-Gly junction.[55] A family of structures has been submitted to the Protein Data Bank under the code 1le1.

The system was simulated with a methyl cap on the C-terminus and charged N-terminus, as in the experiments. Since considerable entrapment was encountered at room temperature, a temperature of 50 °C (323 K) was chosen for the simulations to speed up conformational sampling. This is closer to the experimental melting temperature of 345 K,[55] so that considerable flexibility is expected. A series of eight MC runs was performed, starting from linear extended conformations for up to 800 M configurations. Two simulations successfully found the native state. Figure 4 shows the key computed quantities, total energy, number of backbone interstrand hydrogen bonds, radius of gyration, and the backbone (C, C_α, N) RMSD to the NMR structure as a function of MC steps for one of the trajectories. The first conformer from PDB entry 1le1 was chosen as the reference structure.

In all runs, an immediate relaxation of the system from the extended state is observed, followed by frequent transitions between folded states. After staying trapped in compact coil or β-hairpin conformations for considerable MC time, usually with several backbone hydrogen bonds formed, the system quickly unfolds again to extended conformations. After a short time, another collapse is observed. All runs contain multiple repeats of these folding cycles, as reflected clearly in the gyration radius and number of hydrogen bonds. This behavior is much more dynamic than that observed for the ubiquitin-based β-peptide above, a consequence of the smaller system size and the vicinity to the melting temperature. The collapsed conformations with longer lifetime are β-hairpins and to a lesser extent coiled structures. The turn for the β-structures is always in the middle of the chain, involving the Asn6, Gly7, and Lys8 residues. The native structure as observed in the NMR measurements shows a type I′ β-turn with the Gly at position 3, Glu-Asn-Gly-Lys. This turn is indeed frequently formed during the simulations, often with half of the native β-hairpin formed and only the chain ends frayed. There is one competing β-hairpin that is also highly populated during the runs. It shows a type II′ β-turn with the Gly at position 2, Asn-Gly-Lys-Trp. The backbone interstrand hydrogen bonds for this structure are shifted with respect to the native conformation, and the side chains show very different packing. Other β-conformations are only infrequently observed.

The simulation depicted in Figure 4 finds the native state after 350 M configurations and remains in this state for the rest of the run. The average backbone RMSD to the NMR structure is an extremely low 0.60 ± 0.09 Å over the folded phase, and all native interstrand hydrogen bonds, Ser1(O)−Lys12(N), Thr3(N)−Thr10(O), Thr3(O)−Thr10(N), and Glu5(N)−Lys8(O), are formed except for the Glu5(O)−Lys8(N) pair. Figure 4 shows an overlay plot of a representative structure from this phase with the NMR structure revealing the remarkable similarity. The system has an average energy of −619.2 ± 9.8 kcal/mol and an average number of 3.9 backbone hydrogen bonds over the folded phase. The side-chain structure is also well defined with close resemblance to the NMR data. At this elevated temperature, there is considerable flexibility with occasional flipping of tryptophan residues and high conformational mobility for the lysine side chains. The other trajectory that locates the native state exhibits behavior very similar to that of the first one, finding the native state after 550 M configurations and remaining in this state for the rest of the run.

There are some interesting comparisons for the behavior in the MC runs of trpzip2 with respect to the larger β-hairpin U(1−17)T9D, where all runs led to folded stable β-structures. The trpzip2 simulations were at a 20° higher temperature and were run for 5−10 times as many MC configurations. This was sufficient to unfold locally trapped conformations, and only the native state had a lifetime longer than the simulation length. The six runs that did not sample the native state are not trapped as in the case of U(1−17)T9D, but simply have not found the native basin in their simulation period. The MC progression resembles a random search for the native state with no required intermediate states. For the runs that lead to the native state, final folding is rapid and starts from a completely extended conformation, as can be seen in the gyration radius and RMSD (Figure 4). The transition took only 2 M configurations. The correct turn formed at the chain center and the opposing sides of the β-strand quickly came into contact. All hydrogen bonds formed virtually at the same time, so zipping did not occur.

(55) Cochran, A. G.; Skelton, N. J.; Starovasnik, M. A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 5578−5583.

This is very similar to the folding progression observed for the successful run 1 of U(1−17)T9D.

Since only two of the original eight simulations resulted in a successful fold, a series of eight MC runs at a higher temperature, 70 °C, was also performed. This is just short of the experimental transition temperature of 345 K, so better conformational sampling is expected. Indeed, the simulations exhibit more dynamic behavior with frequent transitions between β-like collapsed states and extended unfolded conformations similar to the simulations at 50 °C, but at a much higher pace. Two of the eight simulations locate the native state (backbone RMSD < 0.8 Å), and the system remains in the native state for 25 or 35 M MC steps before unfolding starts from the chain ends. The side-chain packing is flexible at 70 °C, and the facile unfolding of the native state at this temperature is expected on the basis of the experimental behavior.

Identification of Native States. The success of the present MC approach for predicting polypeptide structures depends on the ability of the method to sample all relevant structures and of the underlying classical energy functions to distinguish non-native from native alternatives. The aforementioned study by Felts et al.[24] gave a reassuring indication that the OPLS-AA potentials are indeed useful in this regard. The present results support the conclusion. The sampling of all relevant parts of phase space appears to have occurred for all of the systems based on the fact that the native state was located in each case. A revealing analysis of the overall energetic landscape for each polypeptide can then be made using simple plots of the total potential energy vs RMSD, as shown in Figure 5. Results at intervals of 0.1 M configurations from all runs were used to make the plot. For a native structure to stand out, most of the lowest-energy structures that are sampled should concentrate in a similar RMSD region.

The U(1−17)T9D peptide shows a strong correlation between energy and RMSD with the native state significantly stabilized with respect to other conformations (Figure 5a); almost all structures with energies below −820 kcal/mol are in the native well. In the case of α-1, the reference RMSD is to a completely folded α-helix. In this case, it is clear that no specific cluster of conformations is strongly preferred in energy; there is a wide RMS range for the lowest-energy structures (Figure 5b). The large cluster in the middle of the plot represents partially helical structures, and the narrow cluster above 6 Å contains β-hairpins. Completely helical conformations at low RMS are relatively high in energy. The landscape is consistent with an essentially unstructured system. Then in Figure 5c, the case of the tryptophan zipper is more similar to that for the U(1−17)T9D peptide. The energetic separation is weaker with a few non-native conformations reaching energies just slightly above the native state. However, the large majority of structures with energies below −630 kcal/mol fall in the native well.

Overall, the native cluster is in the deepest energy band; there is a gradual increase in the energy with rising RMSD in passing to the misfolded hairpins and finally to the extended conformations. This energy landscape can be compared with the results from the study of the same system using similar methodology by Okur et al.,[56] who have employed molecular dynamics with the AMBER ff94 and ff99 force fields and with both the GBSA solvation model and explicit representation of the solvent with TIP3P water. Comparison of the RMSD versus energy graphs reveals striking differences to the present results. In their studies, non-native helical states were found to be highly populated and significantly lower in energy than the native state with both force fields. To remove the helical bias, modified torsional parameters for the Φ and Ψ angles were devised. For trpzip2, this shifted the correctly folded manifold to then be ca. 8 kcal/mol lower in energy than the misfolded structures, which is somewhat smaller than the gap in the present work (Figure 5c). The present results provide no evidence for a need to modify the original OPLS-AA force field.[12] However, it should be noted that we have adopted modified parameters for cysteine,[42] which is not a residue in the present peptides.

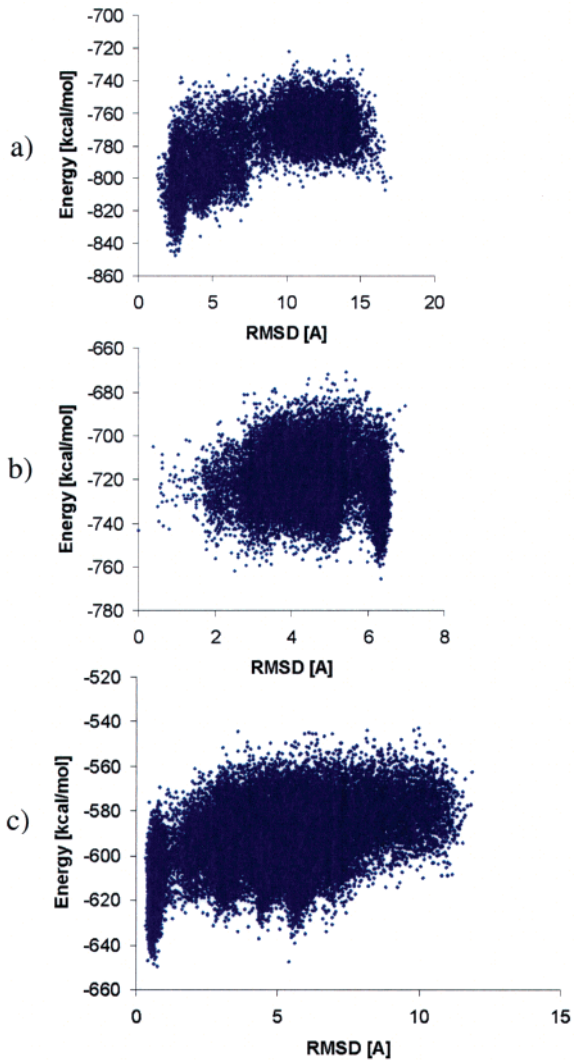

Figure 5. Energy versus RMSD plots for the three polypeptides, (a) U(1−17)T9D, (b) α_1, (c) trpzip2. Snapshots were taken every 105 MC steps, and the results of all runs are shown.

Conclusions

Successful ab initio polypeptide folding is challenging in view of the need to have procedures that provide thorough conformational sampling and correct description of the intramolecular energetics and solvation. In the current work, the combined use of Monte Carlo statistical mechanics with concerted backbone rotations, the OPLS-AA force field, and the GBSA solvation

(56) Okur, A.; Strockbine, B.; Hornak, V.; Simmerling, C. *J. Comput. Chem.* **2003**, *24*, 21−31.

model has been shown to allow successful folding of three polypeptides. The MC simulations started from completely extended conformations and found the correctly folded structures with routine simulation efforts covering 20−800 M configurations. In contrast to some related studies using molecular dynamics,[26,27] the use of significantly elevated temperatures was not found to be necessary; transitions to and between well-folded structures are obtained with the MC methodology at 30−0 °C. The concerted rotations with flexible torsion and bond angles permit efficient consideration of comparatively large conformational changes. The lowest-energy manifolds for the folded structures of the two β-hairpins agree remarkably well with those from the corresponding NMR structures. In the case of the α_1 peptide, the folded α-helical state, which is observed as oligomers in concentrated solution and crystals, is not stable in isolation. The computed preference for random coil structures is in agreement with NMR experiments at low concentration. The fact that native states can be located on high dimensional energy surfaces starting from extended conformations indicates that the present methodology samples well all relevant parts of the conformational space. The OPLS-AA force field with the GBSA solvent model was also found to perform admirably in leading to clear energetic separation of the correctly folded structures from misfolded structures for the two peptides that form β-turns. Optimization of the MC procedures and extension to larger systems are being pursued.

Acknowledgment. Gratitude is expressed to Dr. Julian Tirado-Rives for assistance and to the National Science Foundation (CHE-0130996) and the National Institutes of Health (GM32136) for support of this work.

JA0378862