Parallel Tempering Simulations of HP-36

Chai-Yu Lin,[1] Chin-Kun Hu,[1] and Ulrich H. E. Hansmann[2]*
[1]*Institute of Physics, Academia Sinica, Taipei, Taiwan*
[2]*Department of Physics, Michigan Technological University, Houghton, Michigan*

ABSTRACT **We report results from all-atom Monte Carlo simulations of the 36-residue villin headpiece subdomain HP-36. Protein-solvent interactions are approximated by an implicit solvent model. The parallel tempering is used to overcome the problem of slow convergence in low-temperature protein simulations. Our results show that this technique allows one to sample native-like structures of small proteins and points out the need for improved energy functions. Proteins 2003;52:436–445.** © 2003 Wiley-Liss, Inc.

Key words: protein-folding problem; structure prediction; Monte Carlo simulation; implicit solvent

INTRODUCTION

Evaluating the structure and function of proteins solely from their sequence of amino acids is still a challenge. Attempts to investigate the sequence-structure relationship by means of computer simulations are hampered by the problem that the energy landscape is characterized by a multitude of local minima separated by high-energy barriers. At low temperatures, simple canonical Monte Carlo or molecular dynamics simulations of realistic protein models will not thermalize within a finite amount of available CPU time, and physical quantities cannot be calculated accurately. A second problem is the reliability of protein models. Experimental evidence suggests that the biologically active state of a protein is its global minimum in free energy at room temperature (which for sufficiently large molecules can be approximated by the global minimum in potential energy). However, this is not necessarily true for the available energy functions that only approximate the interactions between atoms within a protein and between the protein and surrounding solvent. For this reason, structure prediction of proteins is limited by the accuracy of the energy functions.

A number of novel techniques that overcome the problem of poor sampling in protein simulations have been developed over the last few years.[1] For instance, generalized ensemble methods[2] proved to be successful in calculating reliable low-temperature estimates of thermodynamic quantities.[3] In the present article, we investigate whether these new and sophisticated algorithms allow in conjunction with present energy functions the structure prediction of small proteins. We go beyond previous work in which similar questions were studied for small peptides such as the pentapeptide met-enkephalin.[4] We have studied for this purpose the 36-residue villin headpiece subdomain HP-36. The structure of this molecule has been resolved by NMR analysis[5] and is shown in Figure 1. As one of the few small proteins that have a well-defined secondary and tertiary structure and can fold autonomously, it is at the same time sufficiently complex and small enough to make simulations feasible. Choice of this protein allows us also to compare our results with NMR data[5] and with related work.[6,7]

Our results rely on an all-atom representation of the molecule; therefore, our work differs from similar studies (see, e.g., Ref. 8) that rely on reduced protein models. The intramolecular interactions are described by the commonly used ECEPP/2 force field[9] and the protein-solvent interactions are approximated by the solvent accessible surface term of Ooi et al.[10] Gas-phase simulations are added for comparison and to separate the effects of intramolecular interactions and of hydration on folding. One of the generalized ensemble techniques, parallel tempering,[11] is used to obtain reliable estimates of thermodynamic quantities over a large temperature range. Quantities such as the average helicity, number of contacts, average energy, specific heat, radius of gyration, and solvent-accessible surface area are calculated. Although our results point out the need for improved energy functions, they also show that existing force fields allow sampling of native-like conformations if solvation effects are taken correctly into account.

METHODS

Our investigation of the folding physics of HP-36 is based on a detailed, all-atom representation of that protein. The interaction between the atoms is described by a standard force field, ECEPP/2,[9] (as implemented in the program package SMMP[12]) and is given by:

$$E_{\mathrm{ECEPP/2}} = E_{\mathrm{C}} + E_{\mathrm{LJ}} + E_{\mathrm{HB}} + E_{\mathrm{tor}}, \tag{1}$$

$$E_{\mathrm{C}} = \sum_{(i,j)} \frac{332 q_i q_j}{\varepsilon r_{ij}}, \tag{2}$$

Grant sponsor: National Science Foundation; Grant number: CHE-9981874; Grant sponsor: National Science Council of the Republic of China (Taiwan); Grant number: 91-2112-M001-056.

C.-Y. Lin's present address is Department of Physics, National Chung Cheng University, Chiayi 612, Taiwan 11529.

C.-K. Hu's e-mail address is huck@phys.sinica.edu.tw.

*Correspondence to: Ulrich H. E. Hansmann, Department of Physics, Michigan Technological University, Houghton, MI 49931-1291. E-mail: hansmann@mtu.edu.

Received 3 September 2002; Accepted 18 November 2002

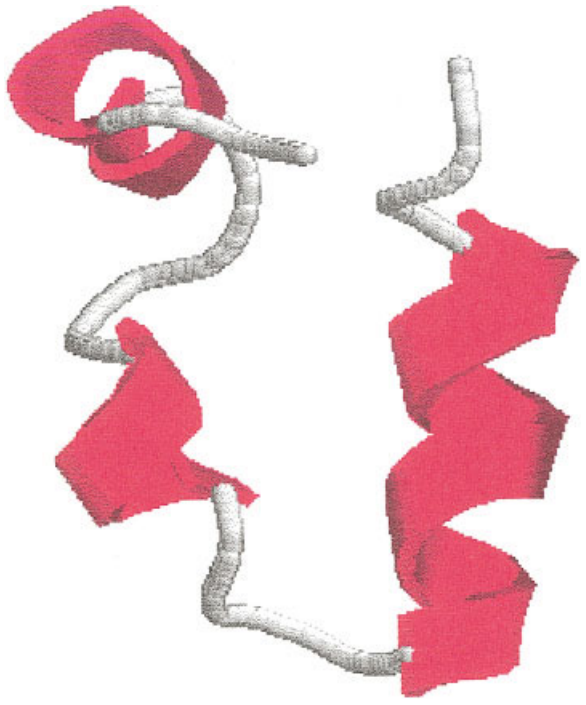

Fig. 1. NMR-derived structure of the 36-residue peptide HP-36 as deposited in the Protein Data Bank (1vii). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

$$E_{LJ} = \sum_{(i,j)} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right), \tag{3}$$

$$E_{\mathrm{HB}} = \sum_{(i,j)} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{6}} \right), \tag{4}$$

$$E_{\mathrm{tor}} = \sum_{l} U_l (1 \pm \cos(n_l \chi_l)). \tag{5}$$

Here, r_{ij} (in Å) is the distance between the atoms i and j, and χ_l is the l-th torsion angle. The protein-water interactions are approximated by a solvent-accessible surface term following the common assumption that the free-energy difference between solvated and unsolvated groups is proportional to the surface area that is exposed to water. Within this approximation, the solvation energy E_{solv} of a protein is given by:

$$E_{\mathrm{solv}} = \sum_{i} \sigma_i A_i. \tag{6}$$

Here A_i is the conformation-dependent solvent-accessible surface area of the i-th atom and σ_i the solvation parameter for the atom i. For the present investigation, we use the solvation parameter set OONS of Ref. 10 that is commonly used together with the ECEPP force field. The potential energy of the solvated molecule is then given by

$$E_{\mathrm{tot}} = E_{\mathrm{ECEPP/2}} + E_{\mathrm{solv}}. \tag{7}$$

Simulations of the solvated molecule are augmented by gas-phase simulations where the solvent-protein interaction term E_{solv} is omitted.

The energy landscape of proteins in such a detailed representation is characterized by a multitude of local minima separated by high-energy barriers. Generalized ensemble methods have been increasingly recognized as a way to overcome the resulting problem of slow convergence in simulations of detailed protein models. One popular example is parallel tempering[11] (also known as replica exchange method or Multiple Markov chains), a technique that was first applied to protein studies in Ref. 13.

In its most common form, one considers in parallel tempering an artificial system that is built up out of N non-interacting replicas of the molecule, each at a different temperature T_i. In addition to standard Monte Carlo or molecular dynamics moves that effect only one copy, parallel tempering introduces a new global update[11]: the exchange of conformations between two copies i and $j = i + 1$ ($i \geq 1$ and $j \leq N$). This replica exchange move is accepted or rejected according to the Metropolis criterion with probability

$$w(C^{\mathrm{old}} \rightarrow C^{\mathrm{new}}) = \min(1, \exp(-\beta_i E(C_j) - \beta_j E(C_i) + \beta_i E(C_i) + \beta_j E(C_j))) \tag{8}$$

Through the exchange of conformations, the Markov chain converges at low temperatures much faster toward the stationary distribution than it does in the case of a regular canonical simulation with only local moves.

The parallel tempering technique does not require Boltzmann weights and can be easily combined with other generalized ensemble techniques.[13] In our study, we have implemented parallel tempering in the simple form described above. For this purpose, we have simulated our molecule on 20 nodes of a cluster of IBM 4-ways 375MHZ SMP Thin Nodes. We have chosen as temperatures $T =$ 1000, 900, 800, 700, 610, 560, 530, 510, 495, 485, 475, 465, 450, 420, 390, 360, 330, 300, 275, and 250 K. On each node, we have performed 150,000 MC sweeps starting from a completely random configuration ("Hot Start"). Each sweep consist of N_{DA} Monte Carlo moves, one for each of the $N_{\mathrm{DA}} = 207$ dihedral angles. A replica exchange move is attempted after each sweep. We display in Figure 2 for a typical replica the "time series" of temperatures and energies that are visited in the course of the simulation. Because of the successive exchange of conformations, the replica moves randomly between high and low temperatures. It is important that the highest temperature T_{Max} is chosen such that any energy barrier can be overcome (see the corresponding "time series" in energy). Only then will the replica exchange move ensure that the molecule thermalizes at all temperatures. For this reason, we have chosen as highest temperature $T_{\mathrm{Max}} = 1000$ K. In hindsight, it is clear that a lower T_{Max} would have been sufficient as long as $T_{\mathrm{Max}} > T_{\mathrm{C}}$ where T_{C} is introduced in the next section.

Monitoring the "time series" of energies for the lowest temperature we find that the system is for $T = 250$ (the

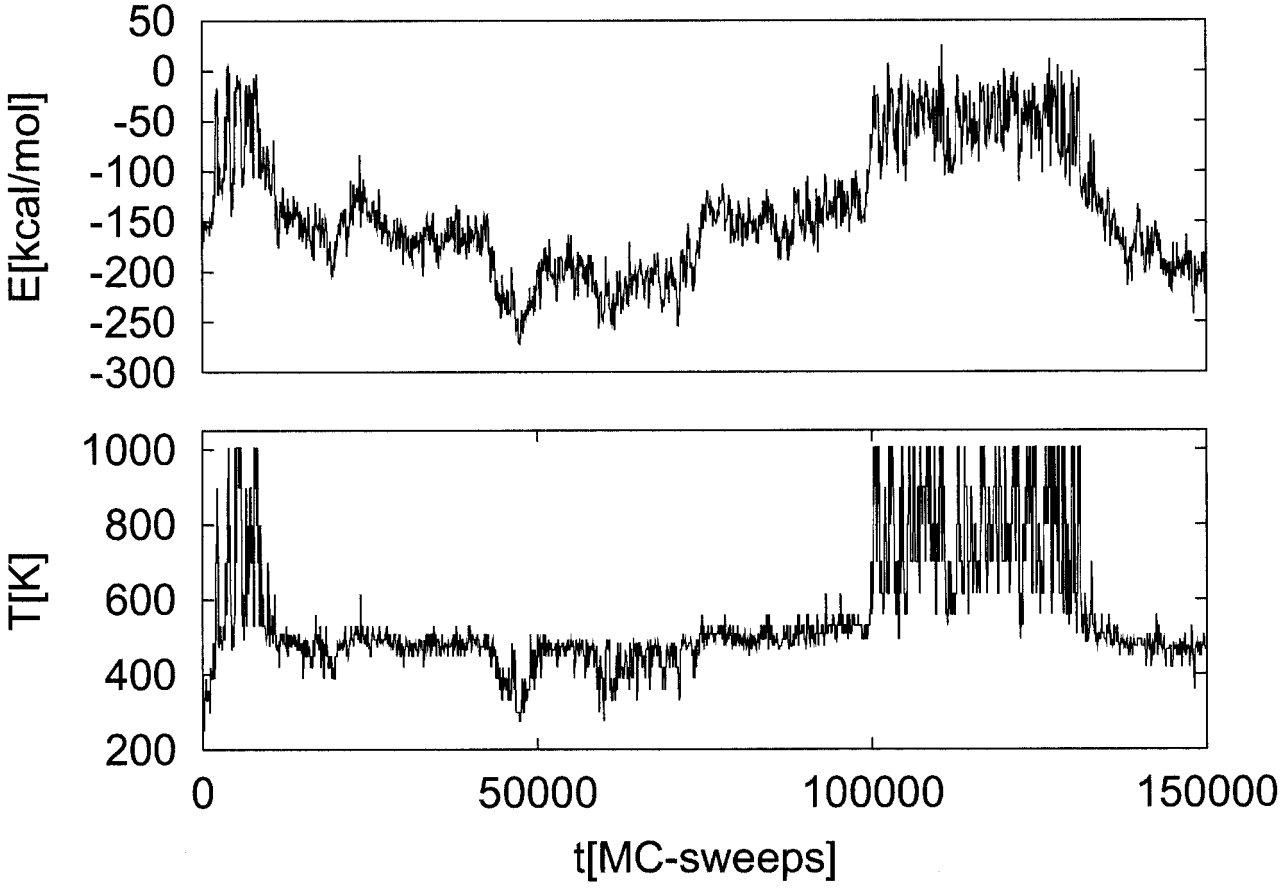

Fig. 2. Movement of a typical replica over the whole temperature range by parallel tempering moves. The corresponding time series in energy is also displayed.

lowest temperature) in equilibrium after 25,000 sweeps. Therefore, we discard this number of sweeps and use for our analysis only the remaining 125,000 MC sweeps. Measurements are taken every 10 Monte Carlo sweeps and written as "time series" in a data file for later analysis. Quantities that are measured include the total E_{tot}, intramolecular energy $E_{\mathrm{ECCEP/2}}$, solvation energy E_{solv}, radius of gyration R_{GY}, solvent-accessible surface area A (as determined by the double cubic lattice method[14]), number of helical residues n_{H}, and total number of contacts n_{TC}. Here, we follow earlier work[15] and define a residue as helical if the pair of backbone dihedral angles (ϕ, φ) takes a value in the range $(-70 \pm 30, -37 \pm 30)$. Likewise, we define two residues as in contact if their C_α-atoms are closer than 8.5 Å. For comparison with experimental data, we have taken the Protein Data Bank structure of HP-36 (PDB code 1vii) and regularized it with the program FANTOM.[16] Contacts that appear both in a given configuration and in this regularized structure are called native contacts n_{NC} and measured by us.

RESULTS AND DISCUSSION

Similar to other generalized ensemble techniques, parallel tempering allows one to evaluate thermodynamic quantities over a range of temperatures. We use this technique to calculate the average total energy $\langle E_{\mathrm{tot}} \rangle$ as a function of temperature both from the simulation of the protein with an OONS solvent (Fig. 3) and in gas phase (Fig. 4). In both plots, we display the specific heat $C(T)$ as an inset. The latter quantity is defined by

$$C(T) = \beta^2(\langle E_{\mathrm{tot}}^2 \rangle - \langle E_{\mathrm{tot}} \rangle)/N$$

where $N = 36$ is the number of residues. Note that $C(T)$ can be interpreted physically only as a true temperature derivative of the partition function if one assumes that the solvation model is temperature independent. This is in general not true, and the specific heat represents, therefore, only the fluctuations in the canonical ensemble of the effective potential of mean force that is used in the simulations.

Although $\langle E_{\mathrm{tot}} \rangle$ decreases gradually with temperature in gas phase, we observe in OONS simulations (Fig. 3) a steep decrease of the average total energy at a "critical" temperature $T_{\mathrm{C}} = 490 \pm 10$ K. Correspondingly, we observe at that temperature a pronounced peak in the specific heat $C(T)$ that is missing in the gas-phase simulation (Fig. 4). The peak in $C(T)$ and the corresponding drop in $\langle E_{\mathrm{tot}} \rangle$ indicate that one observes in OONS simulations a transition between two states that is missing in gas phase. We conjec-

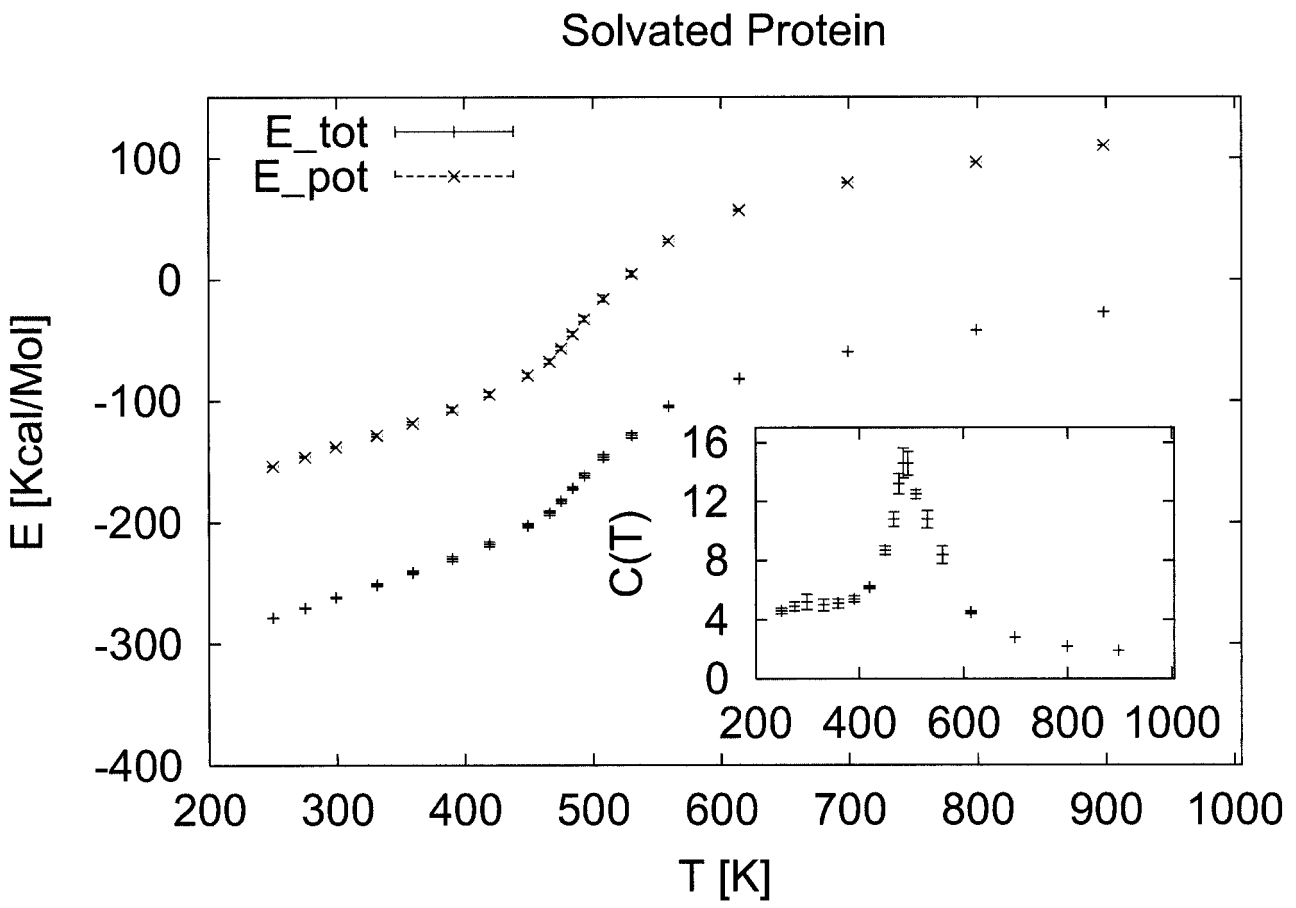

Fig. 3. Average energy $\langle E \rangle(T)$ of HP-36 as a function of temperature. The inlet displays the specific heat $C(T)$. The data were calculated from parallel tempering simulations of the protein using a solvent-accessible surface term to approximate protein-water interactions.

ture that this transition separates the high-temperature region of predominantly unfolded configurations from a phase at temperatures below $T \approx 490$ K where configurations similar to the native one appear. To test the conjecture, we have displayed in Figure 5 the average number of native contacts $\langle n_{\mathrm{NC}} \rangle(T)$ as a function of temperature. This quantity measures the similarity between a protein configuration and the experimentally determined PDB structure (by counting the contacts that appear in both structures). We observe for the solvated protein at the transition temperature $T_{\mathrm{C}} \approx 490$ K a pronounced increase of $\langle n_{\mathrm{NC}}(T) \rangle$. At high temperatures, the number of native contacts is small (and therefore also the resemblance to the native structure), whereas we find at low temperatures (say at $T = 300$ K), configurations that have on average 75% of the native contacts. On the other hand, $\langle n_{\mathrm{NC}}(T) \rangle$ increases only gradually with decreasing temperature in gas-phase simulation. It is even at low temperature substantially lower than found in OONS simulations: on average only 45% of native contacts are formed at $T = 300$ K. The fluctuations of $\langle n_{\mathrm{NC}}(T) \rangle$, defined by

$$\Delta = (\langle n_{\mathrm{NC}}^2 \rangle - \langle n_{\mathrm{NC}} \rangle^2)/N,$$

are displayed in the inset. Although one finds for the OONS simulation a pronounced peak in this quantity at the temperature T_{C} where the specific heat has a maximum, the corresponding curve for the gas-phase simulations is broader and less pronounced. This finding indicates again that the peak in specific heat is indeed correlated with a folding transition in this peptide, which is only observed when the OONS term is added in the simulation. We remark, however, that the transition temperature of $T_{\mathrm{C}} \approx 490$ K in OONS simulations is still unphysiologically high. This finding indicates limitations of our energy function (i.e., the combination of the ECEPP/2 force field with a solvent-accessible surface term and the OONS parameter set to approximate the protein-solvent interaction).

How different are the configurations obtained in gas phase from the ones in OONS simulations? Figure 6 displays the number of helical residues $\langle n_{\mathrm{H}} \rangle(T)$ as function of temperature. Little difference is found at high temperatures; however, below the transition temperature $T \approx 490$ K, the data for both simulations diverge. Now, the helicity grows rapidly with decreasing temperature in the OONS simulation, whereas it stays small in gas phase. Configurations in gas phase and in OONS simulations differ also in their compactness. We display in Figure 7 for HP-36 two quantities that measure the compactness of protein configurations. The main graph is a plot of the average radius of

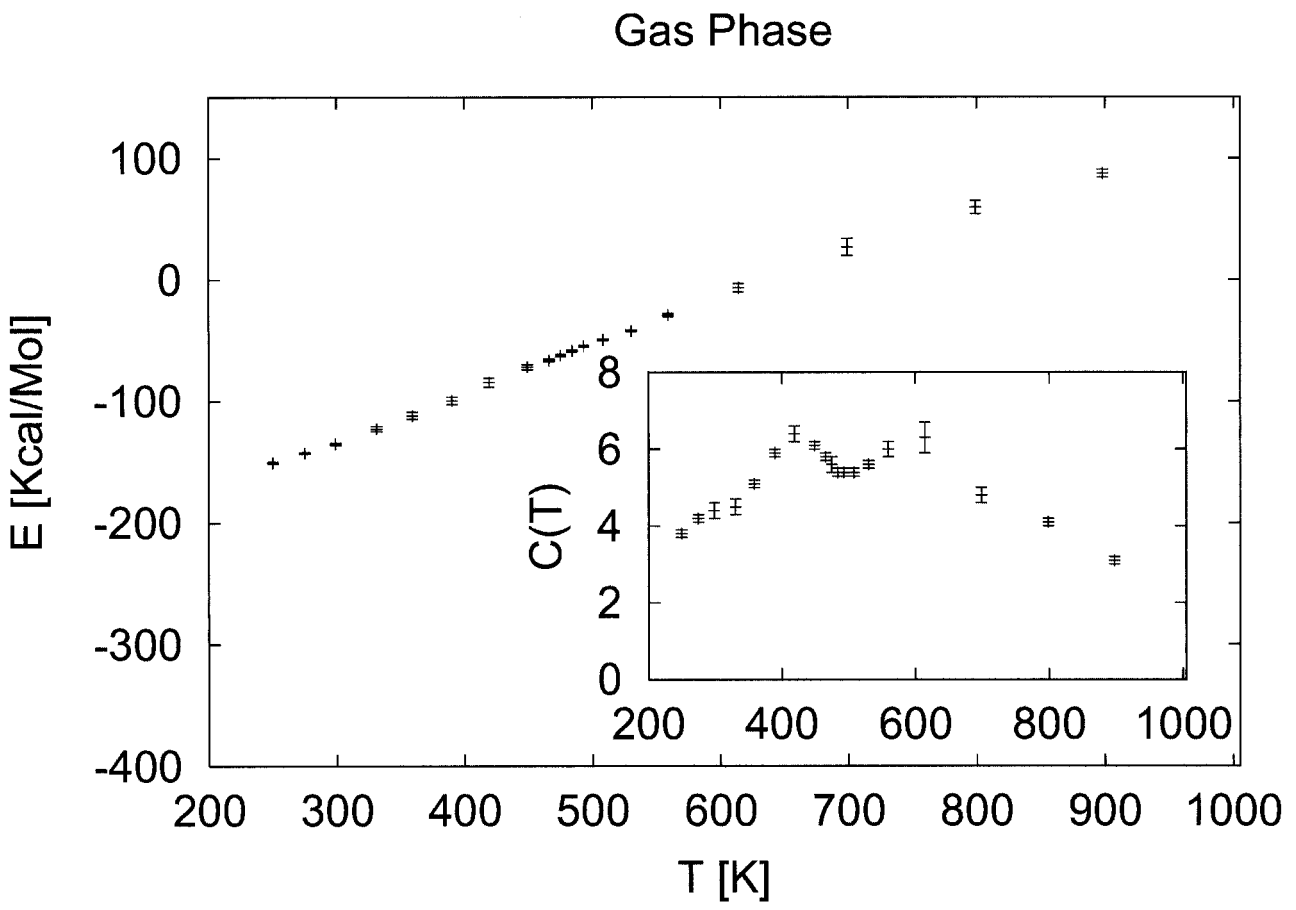

Fig. 4. Average energy $\langle E \rangle(T)$ of HP-36 as a function of temperature. The inlet displays the specific heat $C(T)$. The data were calculated from parallel tempering simulations of the protein in gas phase.

gyration $\langle r_{gy} \rangle(T)$ as a function of temperature. The corresponding values for the total number of contacts $\langle n_{TC}(T) \rangle$ are shown in the inset. Both plots indicate that configurations in gas phase are substantially more compact than the ones in the OONS simulation. For instance, at $T = 300$ K, we find $r_{rgy} = 9.6(1)$ Å in gas phase compared to $r_{rgy} = 12.5(1)$ Å in OONS simulations. Note that even at $T = 1000$ K, the peptide in gas phase has a radius of gyration $r_{gy} = 15.6(1)$ Å and is substantially more compact than in OONS simulation ($r_{gy} = 19.2$ Å). We conjecture that this bias toward compact configurations inhibits the formation of α-helices and that the low-energy states of HP-36 in gas phase are characterized by large density and low helicity.

This conjecture is supported by Figure 8 where we display the lowest energy configuration found in gas-phase simulation. It is essentially a coil structure (with a small helix between residues 23 and 28) and differs strongly from the native one: only 40% native contacts are formed, and the RMSD to the PDB structure is $r_{RMSD} = 7.4$ Å when only backbone atoms are counted and 9.3 Å if all atoms are included. The single helical segment corresponds to the third helix in the PDB structure that stretches from residues 23–32 but is considerably shorter. A number of other residues in the gas-phase structure have pairs of ϕ, φ-angles that are typical for an α-helix. They also appear at positions where the native structure has the other two helices, but they are never in sequence and, therefore, no further helix is formed. Although the helix content of this gas-phase structure is much lower than in the native one, its radius of gyration is with $r_{rgy} = 9.5$ Å comparable with the native one ($r_{rgy} = 9.6$ Å). Minimizing the gas-phase structure further reduces the radius of gyration to $r_{rgy} = 9.3$ Å. Similarly, the solvent-accessible surface area is reduced from 2943 to 2865 Å^2, and the energy from $E_{ECEPP/2} = -168.7$ kcal/mol to $E_{ECEPP/2} = -209.2$ kcal/mol. For comparison, the regularized PDB structure has a solvent-accessible surface area of 2904 Å^2 and an energy of $E_{ECEPP/2} = -176.1$ kcal/mol. Hence, the gas-phase simulation leads to configurations that have comparable or lower energies than the native one but are structurally different. It follows that the native structure is not the global minimum configuration of the ECEPP/2 force field. This was already observed in Ref. 7 where a new global optimization technique was tested for HP-36. It follows that (as expected) gas-phase simulations are not adequate for finding the native structure of proteins of this size. Note also that neither solvent-accessible surface area A nor radius of gyration r_{rgy} allow one to identify the native structure out of an ensemble of low-energy configurations

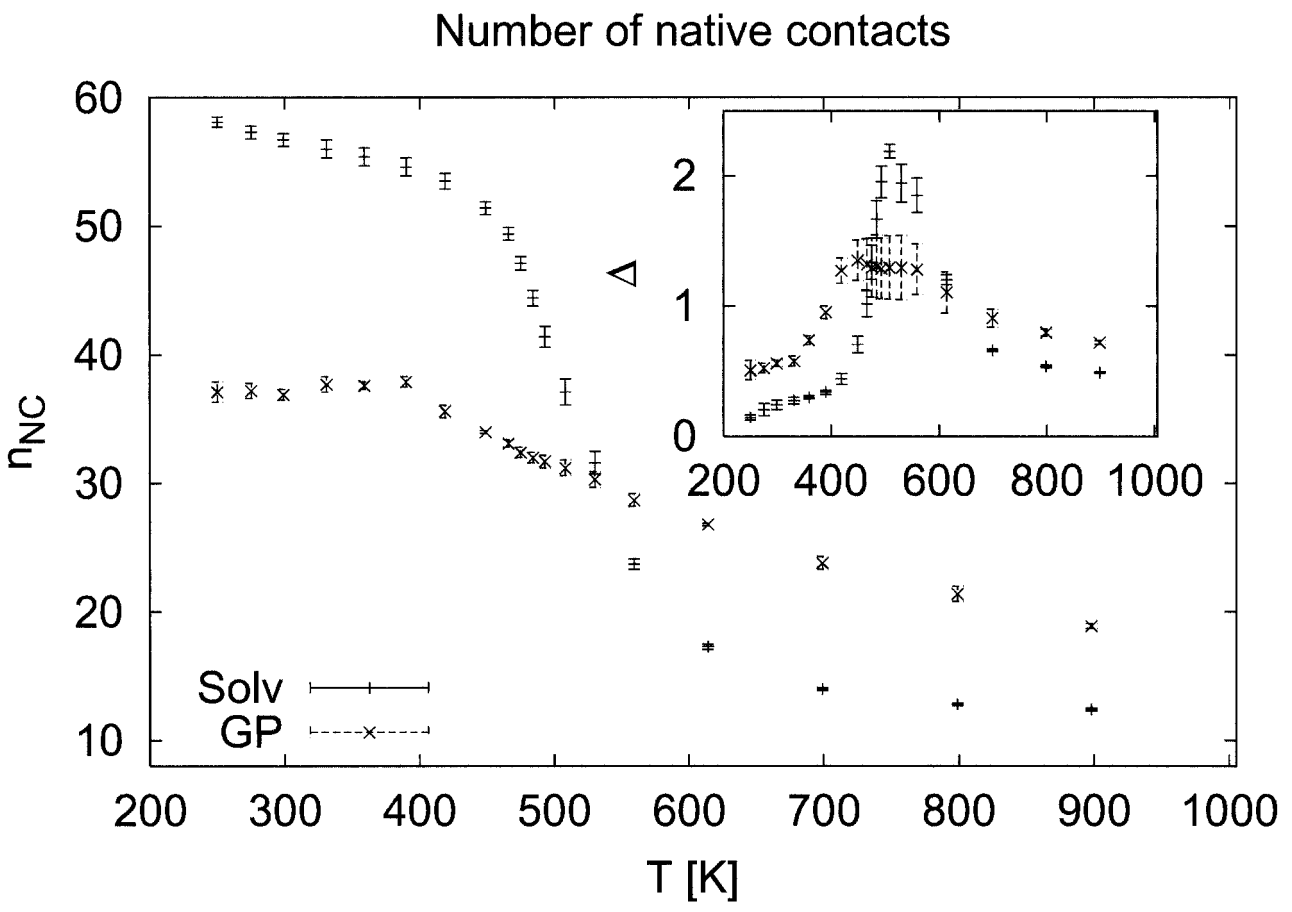

Fig. 5. Average number of native contacts $\langle n_{nc} \rangle(T)$ as a function of temperature. Shown are both results of the solvated protein and in gas phase. The inlet displays the variations of this quantity as a function of temperature.

(i.e., none of these quantities can serve as an "order parameter" for folding).

In Ref. 7, an essentially correct structure was found when a (different) solvent-accessible surface term was added to the energy function. This seems not to be the case in our simulations with the OONS term where we find at $T = 300$ K average values for the radius of gyration of $\langle r_{rgy} \rangle = 12.5(2)$ Å, for the solvent-accessible surface area of $\langle A \rangle = 3688(30)$ Å^2, and for the helicity of $\langle n_H \rangle = 28.3(4)$. These values seem to indicate that the low-temperature configurations in OONS simulations are less dense and have higher helicity than the native structure. However, a careful analysis of our data reveals that two groups of low-energy configurations are found in OONS simulations. The predominant structure (group A) appears at $T = 250$ K with 80% frequency and the lowest-energy configuration of this group is displayed in Figure 9(a). Its elongated structure with ($r_{rgy} = 12.5$ Å) has a large solvent-accessible surface ($A = 3600$ Å^2), and 85% of its residues are part of an α-helix (compared with 54% in the PDB structure). Hence, this structure has little similarity with the native one. Consequently, the structure of Figure 9(a) differs from the regularized PDB structure by root-mean-square deviation (RMSD) of $r_{RMSD} = 7.4$ Å when all backbone atoms are counted (9.3 Å for all atoms).

However, a second group (B) of low-energy configurations is also observed in OONS simulations and appears with 15% frequency at $T = 250$ K. A typical example of these more compact configurations is displayed in Figure 9(b). It has a radius of gyration $r_{rgy} = 9.6$ Å, solvent-accessible surface area $A = 3080$ Å^2, and number of helical residues $n_H = 24$, which are similar to the native structure. All three helices are formed. The first helix stretches from residue 2 to residue 11 and is more elongated than the corresponding one in the PDB structure (residues 4–8). The second helix consists of residues 13–20 and is more elongated than in the PDB structure (residues 15–18). On the other hand, the third helix (residues 26–31) is slightly shorter than in the PDB structure (residues 23–32). The differences in length of the three helices and the shift in the relative position of the first helix (residues 2–11) lead to an RMSD to the PDB-structure of $r_{RMSD} = 5.9$ Å (7.3 Å when counting all atoms). Hence, this structure has an RMSD that is comparable to the structures found in Ref. 7 (an RMSD of $r_{RMSD} = 5.8$ Å when counting backbone atoms) and the 1-μs molecular dynamics simulation of Duan and Kollman[6] (a main-chain RMSD of $r_{RMSD} = 5.7$ Å). Residues Leu2, Phe7, Val10, Phe11, Ala17, Phe18, Lys25, Gln26, Leu29, and Gly34, which compromise or pack against the hydrophobic core of

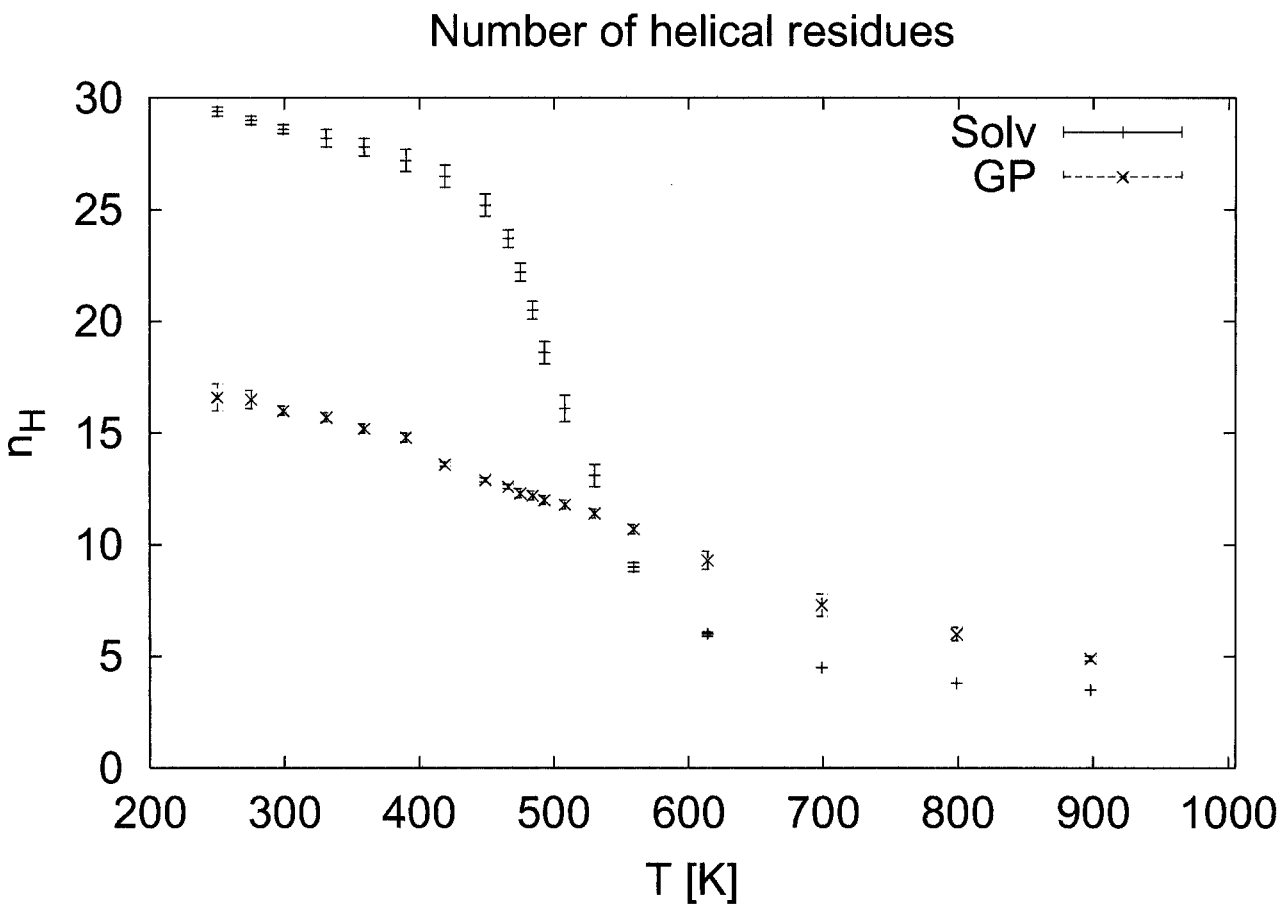

Fig. 6. Average number of helical residues $\langle n_H \rangle(T)$ as function of temperature as obtained from simulations of HP-36 in gas phase and solvated.

HP-36, are >50% solvent inaccessible. However, they are more exposed to the solvent than in the PDB structure where these residues are >70% solvent inaccessible. Our structure has 79% of the native helical content (i.e., 79% of all residues which are part of a helix in the experimental structure are also part of a helix in our structure). In addition, 60% of the native contacts are formed. Both values are only slightly smaller than the results of Ref. 6 (the optimal structure of a 1-μs molecular dynamic folding simulation had 80% of native helical content and 62% of native contacts) and the lowest energy configuration of Ref. 7, which had 90% of helical content and 65% of native contacts.

The above-presented example corroborates that group B consists of low-energy configurations that are similar to the PDB structure. One would expect that configurations of this group (and within this group the native structure) are energetically favored if the energy function were sufficiently accurate. This is not the case in our OONS simulation of HP-36. The average energy of the predominant elongated configurations of group A is at $T = 300$ K $\langle E_{\text{tot}} \rangle = -255(19)$ kcal/mol and differs little from the corresponding value of $\langle E_{\text{tot}} \rangle = -250(16)$ kcal/mol for the more compact and native-like configurations of group B. Both values are comparable to the regularized PDB structure: $E_{\text{tot}} = -253.3$ kcal/mol. We remark that a combination of short canonical simulations at $T = 250$ K with successive minimization leads from the regularized PDB structure to one that is very similar (the main-chain RMSD is 0.9 Å) but has a much lower energy of $E_{\text{tot}} = -306.2$ kcal/mol. For comparison, the lowest energy found for group A structures is $E_{\text{tot}} = -302.9$ kcal/mol (-325.1 kcal/mol after minimization), and this configuration has a backbone RMSD of 7.4 Å to the PDB structure. The lowest energy found for configurations of group B is $E_{\text{tot}} = -299.2$ kcal/mol (-337.8 kcal/mol after minimization), and that structure has a backbone RMSD of 3.7 Å to the PDB structure. Hence, although the combination of ECEPP/2 and OONS solvent term has as global minimum a configuration that is similar to the native one (as was found also with a different parameter set 17 by the global optimization technique of Ref. 7), it does not differentiate folded and other structures at room temperature. Because the configurations of group A have a much larger entropy, native-like configurations are also not the global minimum in free energy. This is in contradiction to the experimental results of Ref. 5. Although we would prefer to have a verification of their results by independent experimental data, we believe that our results show the limitations of our energy function rather than experimental uncertainties.

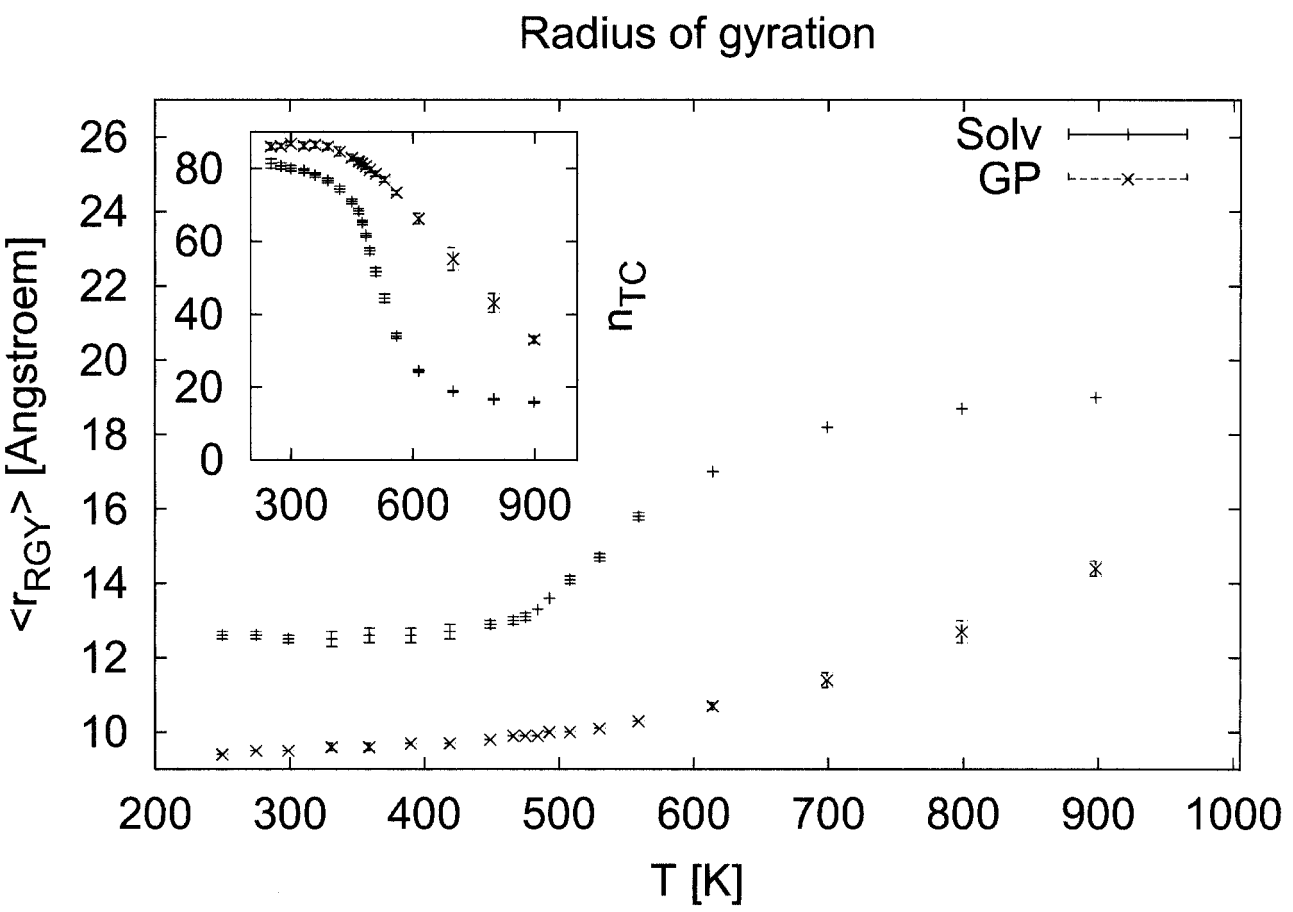

Fig. 7. Average radius of gyration $\langle r_{gy}\rangle(T)$ of HP-36 as a function of temperature for both the solvated protein and in gas phase. The inlet shows the corresponding total numbers of contacts $\langle n_c\rangle(T)$.

The question arises then whether one can distinguish by other means the native structure from misfolded ones in finite temperature OONS simulations. For HP-36, we notice that the PDB structure is characterized through its compactness and high helicity. Hence, we conjecture that in helical proteins the folded structure is the low-energy structure with smallest volume under the constraint that at the same time helicity is maximal (or, more general, the lowest-energy structure that maximizes secondary structure content while minimizing volume). Assuming that this conjecture is true, one can try to define an "order parameter" $O(i)$ that is a function of n_{H} and r_{rgy} and maximal if the configuration i is the native structure. For instance, defining

$$O(i) = \frac{n_{\mathrm{H}}^{\alpha}}{N} e^{\gamma(r_0 - r_{\mathrm{GY}})}, \tag{9}$$

and choosing $\alpha = 2$, $\gamma = 0.5$, and $r_0 = 9.0$ Å, we find for the regularized PDB structure $O_{\mathrm{pdb}} = 0.41$, for the structure of Figure 9(b) (which is the one with maximal value of $O(i)$ found in OONS simulations) a value of $O_{\mathrm{B}} = 0.32$, and for the structure of Figure 9(a) $O_{\mathrm{A}} = 0.12$. For comparison, the maximal value of $O(i)$ that we have found in gas-phase simulations is $O_{\mathrm{GP}} = 0.22$ with the average at $T = 250$ K $\langle O_{\mathrm{GP}}\rangle = 0.1$. Hence, our above defined quantity is able to discriminate between native-like and other configurations for HP-36. However, it is obvious that the definition of $O(i)$ is not universal and will depend on the specific proteins. This limits the usefulness of such an "order parameter" for structure prediction of proteins. Instead, the above conjecture of the native state as one of maximal compactness and secondary structure content may serve better as a heuristic criterion to evaluate low-energy configurations in OONS simulations.

Configurations of group A and B differ also in another quantity: the solvent-accessible surface area. For instance, the elongated configuration of Figure 9(a) (group A) has with $A = 3600$ Å^2 a considerably larger solvent-accessible surface area than the one of Figure 9(b) (group B) with its value $A = 3080$ Å^2. The latter value is closer to the one of the PDB structure where we find $A = 2780$ Å^2. However, the solvent-accessible surface area A alone cannot distinguish native-like conformers from misfolded ones because the mis-folded structures that are found in gas-phase simulations have similar small values. For instance, the configuration that is plotted in Figure 8 has a solvent-accessible surface area of 2865 Å^2. Instead, we observe that the non-polar part A_{NP} of the solvent-accessible

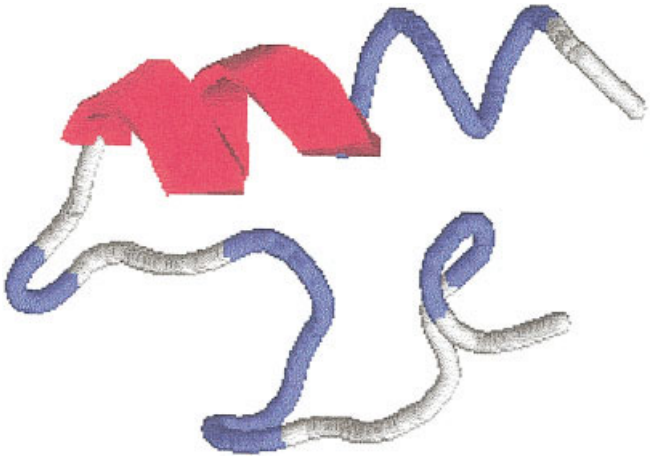

Fig. 8. Typical and/or lowest energy conformation of HP-36 as obtained from a gas-phase simulation.

surface area correlates better with the similarity to the experimentally determined PDB structure. Here, we define this quantity as

$$A_{\mathrm{NP}} = \sum_i \Theta(\sigma(i))A_i \quad \text{with} \quad \Theta(\sigma_i) = \begin{cases} 1 & \text{if } \sigma_i > 0 \\ 0 & \text{if } \sigma_i < 0 \end{cases} \qquad (10)$$

with σ_i the OONS parameter of atom i. With this definition, we find for the PDB structure of Figure 1 $A_{\mathrm{NP}} = 1915$ Å^2. For the misfolded configuration of Figure 9(a), we calculate $A_{\mathrm{NP}} = 2425$ Å^2 and for the native-like configuration of Figure 9(b) $A_{\mathrm{NP}} = 1980$ Å^2. The latter value differs (unlike the solvent-accessible surface area A itself) considerably from the value for configuration of Figure 8 ($A_{\mathrm{NP}} = 2210$ Å^2). The above results for HP-36 indicate that A_{NP} is a suitable criterion for evaluating protein structures in OONS simulations. They further suggest that the OONS parameter set can be improved by slightly raising the values of the parameters σ_i of non-polar atoms [which would increase the energy of structures such as Fig. 9(a) that have larger values of A_{NP}]. We are exploring currently whether such a modified OONS parameter set will allow more efficient structure prediction of HP-36 and similar proteins.

CONCLUSION

We have performed parallel tempering all-atoms simulations of the 36-residue villin headpiece subdomain HP-36. Our results show that this technique allows one to overcome the multiple minima problem in simulations of small proteins and to sample native-like configurations provided that the energy function is sufficiently accurate. Although no native-like configurations are found by us in gas-phase simulations, configurations similar to the PDB structure appear at room temperature in simulations with an additional solvent-accessible surface term (OONS). However, such configurations are found with only 20% frequency and cannot be distinguished by their energies from that of the predominant misfolded structures. Hence, the combination of ECEPP/2 force field with a solvent-accessible surface term and the OONS parameter set is for our protein still not accurate enough for prediction of the native state. Furthermore, as with the global optimization technique of Ref. 7, which relied on a different set of solvation parameters,[17] our structure predictions of HP-36 are limited to an RMSD of ≈6 Å. This finding illustrates the need for improved force fields and better solvent representations. Our results suggest modifications of the OONS parameter set that may lead to a higher accuracy of

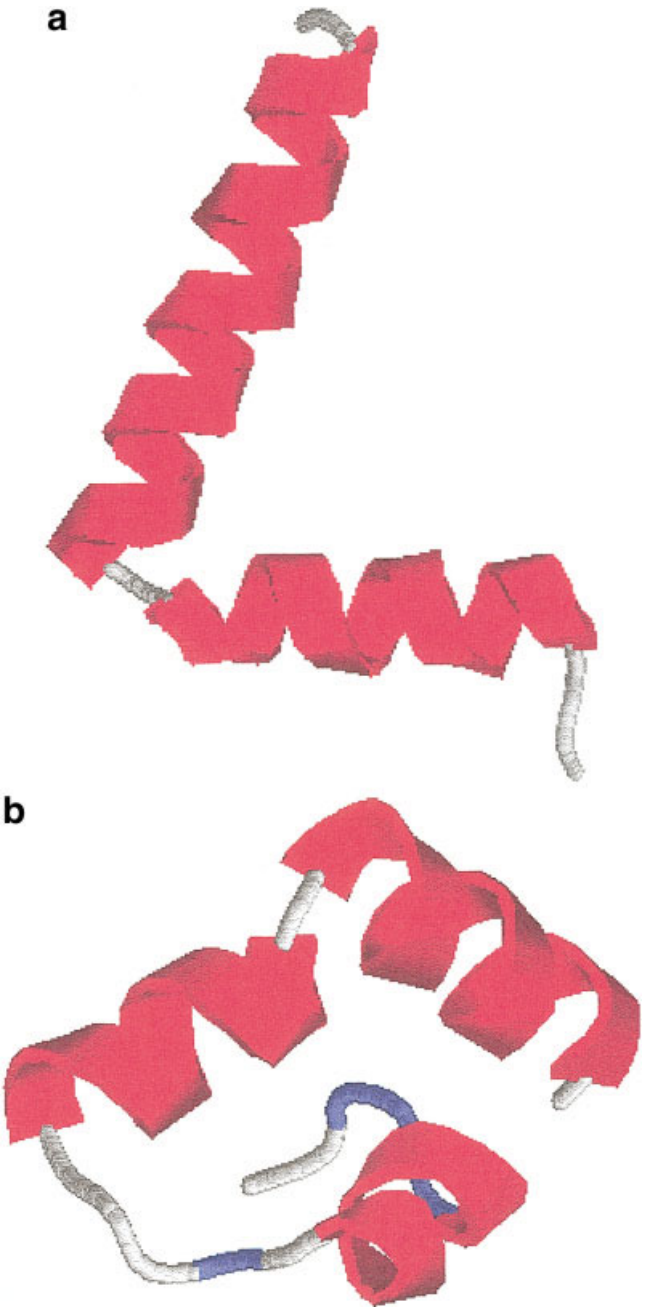

Fig. 9. Low-energy structures of HP-36 as obtained by a simulation with a solvent-accessible surface term. The configuration of (**a**) appears at $T = 250$ K with 80% frequency, whereas the native-like structure of (**b**) is only observed in 15% of the measurements.

the energy function and could increase its usefulness for structure predictions of proteins.

ACKNOWLEDGMENTS

U.H. acknowledges support by research grant CHE-9981874 of the National Science Foundation. We thank the computer center of Academica Sinica (Taipei, Taiwan) for CPU time on their IBM cluster. Part of this work was done while U.H. was visiting the institute of Physics at Academia Sinica (Taipei, Taiwan). He thanks C.K. Hu and the institute for kind hospitality.

REFERENCES

1. Hansmann UHE, Okamoto Y. New Monte Carlo algorithms for protein folding. Curr Opin Struct Biol 1999;9:177–184.
2. Hansmann UHE, Okamoto Y. The generalized-ensemble approach for protein folding simulations. In: Stauffer D, editor. Annual reviews in computational physics. Singapore: World Scientific; 1998.
3. Hansmann UHE, Okamoto Y. Prediction of peptide conformation by multicanonical algorithm: a new approach to the multiple-minima problem. J Comp Chem 1993;14:1333–1338.
4. Klepeis JL, Floudas CA. Comparative study of global minimum energy configurations of hydrated peptides. J Comp Chem 1999;20: 636–654.
5. McKnight CJ, Doehring DS, Matsudaria PT, Kim PS. Thermostable 35-residue subdomain within villin headpiece. J Mol Biol 1996;260:126–134.
6. Duan Y, Kollman PA. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science 1998;282:740–744.
7. Hansmann UHE, Wille L. Global optimization by energy landscape paving. Phys Rev Lett 2002;88:068105(1:4).
8. Favrin G, Irbäck A, Wallin S. Folding of a small helical protein using hydrogen bonds and hydrophobicity forces. Proteins 2002;47: 99–105.
9. Sippl MJ, Némethy G, Scheraga HA. Intermolecular potentials from crystal data. 6. Determination of empirical potentials for O—H · · · O═C hydrogen bonds from packing configurations. J Phys Chem 1994;88:6231–6233 and references therein.
10. Ooi T, Obatake M, Nemethy G, Scheraga HA. Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. Proc Natl Acad Sci USA 1987;84:3086–3090.
11. Hukushima K, Nemoto K. Exchange Monte Carlo method and applications to spin glass simulations. J Phys Soc (Japan) 1996;65: 1604–1608. Geyer GJ, Thompson EA. Annealing Markov Chain Monte Carlo with applications to ancestral inference. J Am Stat Assn 1995;90(431):909–920.
12. Eisenmenger F, Hansmann UHE, Hayryan S, Hu CK. [SMMP] a modern package for simulation of proteins. Comp Phys Comm 2001;138:191–212.
13. Hansmann UHE. Parallel tempering algorithm for conformational studies of biological molecules. Chem Phys Lett 1997;281: 140–150.
14. Eisenhaber F, Lijnzaad P, Argos P, Sander C, Scharf M. The double cubic lattice method: efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. J Comp Chem 1995;16:273–284.
15. Okamoto Y, Hansmann UHE. Thermodynamics of helix-coil transitions studied by multicanonical algorithms. J Phys Chem 1995; 99:11276–11287.
16. Schaumann T, Braun W, Wuthrich K. The program fantom for energy refinement of polypeptides and proteins using a Newton-Raphson minimizer in torsion space. Biopolymers 1990;29:679–694.
17. Wesson M, Eisenberg D. Atomic solvation parameters applied to molecular dynamics of proteins in solution. Protein Sci 1992;1:227–235.