Simulations of a small protein in a specifically designed generalized ensemble

Ulrich H. E. Hansmann*
Department of Physics, Michigan Technological University, Houghton, Michigan 49931-1295, USA
(Received 12 February 2004; revised manuscript received 24 March 2004; published 15 July 2004)

We describe a variant of the generalized-ensemble approach that allows faster simulations for special classes of proteins. We test this technique for an all-atom model of the 36-residue protein HP-36. The dependence of various thermodynamic quantities on small modifications of the solvent representation is explored. Configurations with a root-mean square deviation of less than 4 Å to the experimentally determined structure are observed.

DOI: 10.1103/PhysRevE.70.012902 PACS number(s): 87.15.Aa, 87.15.Cc, 87.15.He

A fundamental problem in molecular biophysics is the relation between the sequence of amino acids in the protein chain and its three-dimensional (3D) shape and function. Computer experiments offer a way to study folding of proteins *in silico*, but are hampered by numerical difficulties [1]. For instance, all-atom models of proteins are characterized by a rough energy landscape, and the resulting slowing down in the search of the conformational space limits the size of proteins that can be studied in low-temperature simulations.

One way to alleviate this multiple-minimum problem is by means of generalized-ensemble algorithms [2]. An example is the technique introduced in this article that promises a much faster sampling for an important class of proteins—namely, such that are built up solely out of α-helices (a fact that can be determined experimentally more easily than the 3D structure). Generalizations of this idea to other classes of proteins are easy to envision.

It is a common practice to test a new algorithm for small and simple molecules such as the pentapeptide Met-enkephalin [3]. While simulations of these molecules are simple (the ground state and folding physics of Met-enkephalin has been successfully studied at room temperature by simple canonical simulations [4]), they can be misleading as the complexity of the problem is different for larger molecules. For this reason, it is important to test the performance of novel simulation algorithms for a sufficiently large and complex molecule. Here, we have chosen the 36 residue villin headpiece subdomain HP-36. As one of the few small proteins that have a well-defined secondary and tertiary structure and can fold autonomously [5], it is sufficiently complex and with 596 atoms large enough that numerical simulations become indeed a challenge [6].

Proteins are only marginally stable. The free-energy difference between the biological state and denatured states is at room temperature only $\approx 10-20$ kcal/mol. However, this small free energy difference results from cancellations of large (both energetic and entropic) terms. It follows that the accuracy of the energy function is another limiting factor in protein simulations. Of special importance here is the protein-water interaction as constraints in available computer time often require the use of implicit solvent models. Unlike small molecules such as Met-enkephalin that do not form a hydrophobic core, HP-36 is well suited to probe the influence of the chosen solvent approximation. We have conjectured in earlier work [7] that the accuracy of structure predictions can be increased for this peptide through modification of our implicit solvent. The test of this conjecture is the second objective of the present HP-36 simulations in our new generalized ensemble.

Our simulations of HP-36 rely on a standard force field ECEPP/3 [8] (as implemented in the program package SMMP [9]), where the intramolecular interactions are given by

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

$$E_C = \sum_{(i,j)} \frac{332 q_i q_j}{\epsilon r_{ij}}, \tag{2}$$

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

$$E_{HB} = \sum_{(i,j)} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right), \tag{4}$$

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

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

$$E_{solv} = \sum_i \sigma_i A_i. \tag{6}$$

Here, A_i is the solvent-accessible surface area and σ_i the solvation parameter of the ith atom. For the present investigation we use the parameter set OONS of Ref. [10].

The energy landscape of proteins in such a detailed representation is characterized by a multitude of local minima separated by high-energy barriers. One way of overcoming the resulting problem of slow convergence is generalized-ensemble simulations first introduced to protein science in Ref. [11]. These techniques rely on simulations in an artificial ensemble designed in such a way that a Monte Carlo or molecular dynamics simulation will lead to a uniform distribution of a prechosen quantity. For instance, in multicanonical sampling [12] the weight $w(E)$ is chosen such that the distribution of energies $P(E)$ is given by

*Electronic address: hansmann@mtu.edu

 ©2004 The American Physical Society

$$P(E) \propto n(E)w(E) = \text{const}, \tag{7}$$

where $n(E)$ is the spectral density. A free random walk in the energy space is performed that allows the simulation to escape from any local minimum. From this simulation one can calculate the thermodynamic average of any physical quantity A by reweighting [13]:

$$\langle \mathcal{A} \rangle_T = \frac{\int dx\ \mathcal{A}(x) w^{-1}(E(x)) e^{-E(x)/k_B T}}{\int dx\ w^{-1}(E(x)) e^{-E(x)/k_B T}}. \tag{8}$$

Here, x stands for configurations and k_B is the Boltzmann constant.

While multicanonical sampling has been successfully applied to polypeptides of up to $\approx$35 residues [14], it is not obvious that the method will succeed for the more important case of molecules that consists of 50–200 amino acids (the size of stable domains in proteins). This is because the computational effort increases in multicanonical simulations with the number of residues as $\approx N^4$ [15]. In general, the computational effort for generalized-ensemble algorithms scales as $\propto X^2$ where X is the variable in which one wants a flat distribution. This is because an unbiased 1D random walk in the ensemble coordinate is generated. In the multicanonical algorithm the coordinate is the potential energy $X=E$. Since $E \propto N^2$, the scaling relation for multicanonical simulations is recovered. Hence, a better efficiency may be obtained by choosing a more appropriate ensemble coordinate than the energy.

The problem is to find such a coordinate as there exists no single "order parameter" for folding. While it is possible to define an order parameter by means of the root-mean-square deviation (rmsd) or the number of native contacts if the native state of the protein is known, such an approach fails for the structure prediction of unknown proteins. However, for special classes of proteins one can often define "order parameters" that that do not require *a priori* knowledge of the native structure. One example is helical proteins—i.e., proteins that are built up solely out of α-helices where the helicity (the number n_H of residues that are part of an α-helix) is a natural choice for distingushing between low-energy conformers. One can now devise a generalized-ensemble algorithm that leads to a (two-dimensional) uniform probability distribution in energy and helicity. However, we found in preliminary runs that a better convergence is obtained by constructing a new (one-dimensional) ensemble that is based on a combination of energy $E_{tot}=E_{ECEPP/3}+E_{solv}$ and helicity n_H as ensemble coordinate:

$$q = \sqrt{E_{tot}^2 + c n_H^2}. \tag{9}$$

It is obvious that the reweighting technique allows one again to calculate thermodynamic averages over a large range of temperatures from a single simulation:

$$\langle \mathcal{A} \rangle_T = \frac{\int dx\ \mathcal{A}(x) w^{-1}(q(x)) e^{-E(x)/k_B T}}{\int dx\ w^{-1}(q(x)) e^{-E(x)/k_B T}}. \tag{10}$$

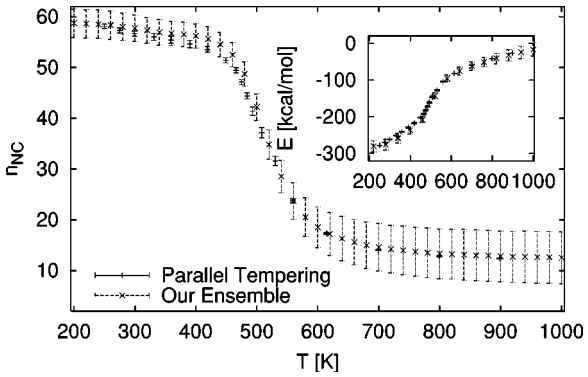

FIG. 1. Average number of native contacts $\langle n_{NC} \rangle(T)$ as a function of temperature. Shown are both results of the parallel tempering simulation of Ref. [7] and such obtained by our approach. The inset displays the corresponding values of the total energy $\langle E_{tot}(T) \rangle$.

As in other generalized-ensembles—and unlike in canonical simulations—the weights are not *a priori* known in simulations in this new ensemble. Instead one has to determine estimators. In the present study, 200 000 sweeps were needed for the calculation of the weights by means of the iterative procedures described in Ref. [16]. All thermodynamic quantities are estimated from a production run of 400 000 Monte Carlo sweeps that followed 10 000 sweeps for equilibrization. The simulations start from completely random initial conformations (hot start) and one Monte Carlo sweep updates every torsion angle of the peptide once. At the end of every tenth sweep, the total energy E_{tot}, the ECEPP/3 energy $E_{ECEPP/3}$, the solvation energy E_{solv}, the corresponding radius of gyration r_{gy}, and the number n_H of helical (sheet) residues are written to a file and stored for later analysis.

While HP-36 is extremely difficult to study in regular canonical simulations [6], we have shown in earlier work [7,17] that the thermodynamic and folding of this peptide is accessible to advanced methods such as parallel tempering [18,19]. Related work can be also found in Ref. [20]. We first try to reproduce these earlier results with our specifically designed ensemble. For this purpose, we display in Fig. 1 the average number of native contacts $\langle n_{NC} \rangle(T)$ as a function of temperatures. 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 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). Likewise, the total energy $\langle E_{tot}(T) \rangle$ is shown in the inlet. Both results from simulations with our new ensemble and from the parallel tempering runs of Ref. [7] are presented in Fig. 1. The latter results rely on a total statistics of 3 000 000 Monte Carlo sweeps—i.e., more than 7 times larger than the statistics in the simulation with our new ensembles. We observe that values of both $\langle n_{NC} \rangle(T)$ and $\langle E_{tot}(T) \rangle$ agree with each other within the error bars for both sets of simulations. Hence, simulations in our new ensemble reproduce successfully the numerical results of earlier work obtained by a different method. This demonstrates that our technique is indeed suitable for simulations of

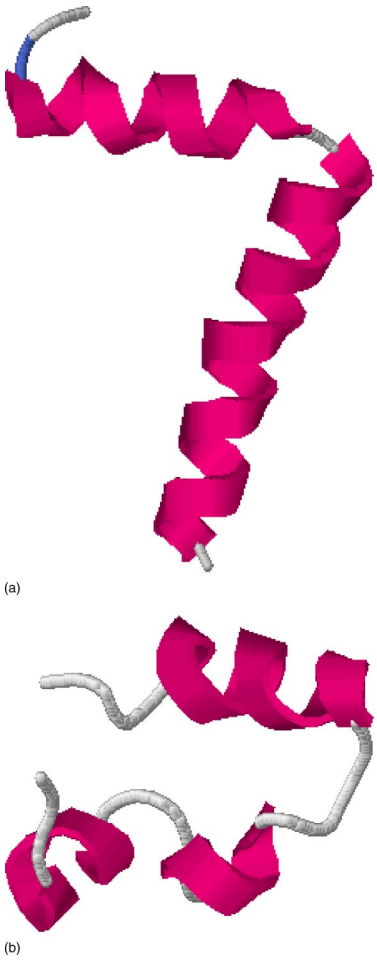

FIG. 2. Low-energy structure (a) of HP-36 as obtained by a simulation with the OONS solvent-accessible surface term. This configuration appears at $T=300$ K with 90% frequency. The remaining 10% resemble the PDB structure (1vii) shown in (b).

helical peptides and proteins with 30–40 residues. Since for this class of molecules our new approach is approximately one order of magnitude computationally more efficient than other generalized-ensemble methods, we plan now to apply it to larger helical molecules such as the B domain of protein A [21].

As already observed in Ref. [7], the native structure as deposited in the PDB is *not* the dominant structure at $T=300$ K in simulations with our energy function but appears at this temperature only with a frequency of about 10%. Instead 90% of the configurations resemble at this temperature the structure shown in Fig. 2(a). For comparison, we show in Fig. 2(b) also the PDB structure of HP-36 (1vii). Both types of configurations have at room temperature similar average energies [7]. Hence, without *a priori* knowledge of the experimental structure it is not possible to identify the native structure as it is in our simulations not the global free energy minimum at this temperature. This is in contradiction to the experimental results of Ref. [5] and indicates severe limitations of our energy function. It was conjectured in Ref. [7] that the low frequency of nativelike configurations is due to the poor approximation of the protein-solvent interaction by the solvent-accessible surface term of Eq. (6). It was further

TABLE I. Original OONS parameter set and its modification M-OONS.

Atom type	OONS	M-OONS
C aliphatic	0.008	0.028
C carboxyl,carbonyl	0.427	0.447
C aromatic	−0.008	0.012
N	−0.132	−0.132
O carboxyl,carbonyl	−0.038	−0.038
O hydroxyl	−0.172	−0.172
O charged	−0.038	−0.038
S	−0.021	−0.021

suggested that this approximation can be improved by enhancing the weight of nonpolar atoms in the OONS parametrization [10] used in our simulations. Since our new ensemble allows for much faster simulations than previously possible, we are now in a position to probe this earlier conjecture of Ref. [7]. As we do not intend to determine the "optimal" parameter set, we restrict ourselves to a single (and not optimized) variation of the OONS set. For this modified parameter set, dubbed M-OONS, the parameters for nonpolar atoms are raised by an arbitrary small value $\Delta\sigma_i=0.02$: $\tilde{\sigma}_i=\sigma_i+\Delta\sigma_i$. Values of σ_i for the modified parameter set M-OONS and the original OONS set are listed in Table I.

M-OONS is used in a simulation of HP-36 with the generalized ensemble given by Eq. (9). As an example of our results we display in Fig. 3 the solvent-accessible volume as a function of temperature. This quantity is a measure for the compactness of configurations. Data for both the original parameter set OONS and the modified set M-OONS are shown. The corresponding values for the helicity are drawn in the inset. The differences in volume and helicity are small at high temperatures. However, the two solvation parameter sets lead to very different results at low temperatures: both volume and helicity are lower for the modified set M-OONS than for the original OONS set. At $T=300$ K, the average solvent accessible volume of HP-36 configurations in the

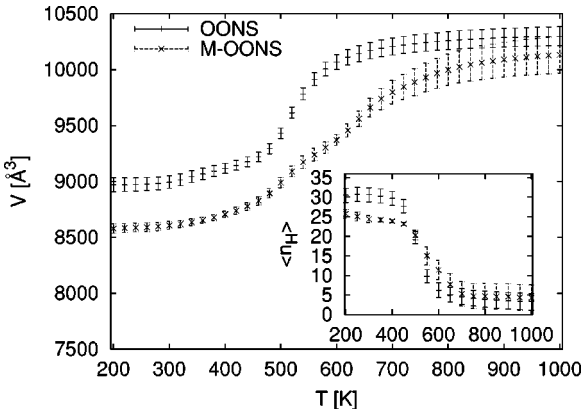

FIG. 3. Average solvent accessible volume $\langle V\rangle(T)$ as a function of temperature for both the OONS parameter set our modified M-OONS set. The inset displays the corresponding values for the average helicity $\langle n_H\rangle(T)$.

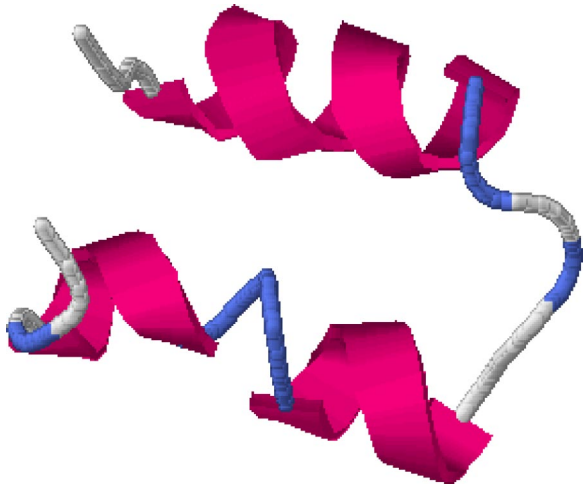

FIG. 4. Low-energy structure of HP-36 as obtained by a simulation with the modified solvent-accessible surface term M-OONS. This configuration appears at $T=300$ K with 99% frequency.

OONS simulation is $\langle V^{OONS}\rangle = 8993(56)\text{Å}^3$ and the average helicity $\langle n_H^{OONS}\rangle = 0.85(5)$, indicating a high frequency of extended and almost completely helical conformations as the one displayed in Fig. 2(a). Simulations relying on the modified set M-OONS lead at room temperature to more compact $[\langle V^{M\text{-}OONS}\rangle = 8605(35)\text{Å}^3]$ configurations with reduced helicity $[\langle n_H^{M\text{-}OONS}\rangle = 0.68(3)]$, and 99% of the observed configurations resemble now the one shown in Fig. 4. The high similarity of this configuration to the PDB structure reflected in a main-chain rmsd of only 3.8 Å (5.0 Å if all heavy atoms are counted), which is much smaller than the corresponding rmsd values (8.1 Å for main-chain atoms, 9.2 Å for all heavy atoms) for the configuration of Fig. 2(a), the dominant structure in simulations with the original OONS parameter set.

The remarkable improvement in the accuracy of structure "predictions" to rmsd values below 4 Å over previous work [6,7,17] that was restricted to rmsd values of 5.8 Å indicates that the OONS parameter set underestimates the hydrophobic effect [16]. Already a slight increase in the parameters of nonpolar atoms (decreasing the helix propensity and increasing alignment of hydrophobic residues) leads at room temperature to results that are more comparable with the experimental structure than the original OONS solvent. We remark that the modified parameter set M-OONS leads also in simulations of the human parathyroid hormone fragment PTH(1-34) to configurations that are closer to the experimentally found structure [22]. More simulations are necessary to determine the optimal solvent parameter set and to probe whether the above result is restricted to helical proteins or is more general. However, such an investigation goes beyond the scope of this Brief Report.

In summary, we have introduced a generalized ensemble that allows efficient simulations of proteins with only α-helices as secondary structure elements. While less general than other search algorithms, the gain in efficiency could open a way to an improved understanding of the folding process for these proteins. As an application we have simulated the 36-residue protein HP-36 and demonstrated that the accuracy of our implicit solvent can be improved by increasing the weight of no-polar atoms. Configurations within 4 Å to the experimentally determined structure are observed.

Support by a research grant from the National Institutes of Health No. (GM62838) is gratefully acknowledged. Part of this article was written while visiting the University of Central Florida in Orlando, FL. I thank the UCF Physics Department for kind hospitality.

[1] U. H. E. Hansmann, Comput. Sci. Eng. **5**, 64 (2003).
[2] U. H. E. Hansmann and Y. Okamoto, in *Annual Reviews in Computational Physics VI*, edited by D. Stauffer (World Scientific, Singapore, 1998), p. 129.
[3] B. A. Berg, Phys. Rev. Lett. **90**, 180601 (2003).
[4] U. H. E. Hansmann and J. N. Onuchic, J. Chem. Phys. **115**, 1601 (2001).
[5] C. J. McKnight *et al.*, J. Mol. Biol. **260**, 126 (1996).
[6] Y. Duan and P. A. Kollman, Science **282**, 740 (1998).
[7] C.-Y. Lin, C.-K. Hu, and U. H. E. Hansmann, Proteins: Struct., Funct., Genet. **52**, 436 (2003).
[8] G. Némethy *et al.*, J. Phys. Chem. **96**, 6472 (1992).
[9] F. Eisenmenger *et al.*, Comput. Phys. Commun. **138**, 192 (2001).
[10] T. Ooi *et al.*, Proc. Natl. Acad. Sci. U.S.A. **8**, 3086 (1987).
[11] U. H. E. Hansmann and Y. Okamoto, J. Comput. Chem. **14**, 1333 (1993).
[12] B. A. Berg and T. Neuhaus, Phys. Lett. B **267**, 249 (1991).
[13] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **61**, 2635 (1988); **63**, 1658(E) (1989), and references given in the erratum.
[14] U. H. E. Hansmann, J. Chem. Phys. **120**, 417 (2004).
[15] U. H. E. Hansmann and Y. Okamoto, J. Chem. Phys. **110**, 1267 (1999); **111**, 1339(E) (1999).
[16] Y. Peng and U. H. E. Hansmann, Biophys. J. **82**, 3269 (2002).
[17] U. H. E. Hansmann and L. T. Wille, Phys. Rev. Lett. **88**, 068105 (2002).
[18] K. Hukushima and K. Nemoto, J. Phys. Soc. Jpn. **65**, 1604 (1996); G. J. Geyer, Stat. Sci. **7**, 437 (1992).
[19] U. H. E. Hansmann, Chem. Phys. Lett. **281**, 140 (1997).
[20] M.-Y. Shen and K. F. Freed, Proteins **49**, 439 (2002).
[21] H. Gouda *et al.*, Biochemistry, **31**, 9665 (1992).
[22] U.H.E.Hansmann (unpublished).