ABSTRACT
Motivation: Conventional Monte Carlo and molecular dynamics simulations of proteins in the canonical ensemble are of little use, because they tend to get trapped in states of energy local minima at low temperatures. One way to surmount this difficulty is to use a non-Boltzmann sampling method in which configurations are sampled upon a general weighting function instead of the conventional Boltzmann weighting function. The multiensemble sampling (MES) method is a non-Boltzmann sampling method that was originally developed to estimate free energy differences between systems with different potential energies and/or at different thermodynamic states. The method has yet not been applied to studies of complex molecular systems such as proteins.

Results: MES Monte Carlo simulations of small proteins have been carried out using a united-residue force field. The proteins at several temperatures from the unfolded to the folded states were simulated in a single MC run at a time and their equilibrium thermodynamic properties were calculated correctly. The distributions of sampled conformations clearly indicate that, when going through states of energy local minima, the MES simulation did not get trapped in them but escaped from them so quickly that all the relevant parts of conformation space could be sampled properly. A two-step folding process consisting of a collapse transition followed by a folding transition is observed. This study demonstrates that the use of MES alleviates the multiple-minima problem greatly.

Availability: Available on request from the authors
Contact: khan@pcu.ac.kr
Supplementary information: A FORTRAN90 code of MES algorithm for MC simulation and its sample input have been deposited as Supplementary data available at Bioinformatics online.

1 INTRODUCTION
The computer experiment employing Monte Carlo (MC) method or molecular dynamics (MD) method is now a popular tool in science. Despite the recent advances of computing power, simulations of large complicated systems such as spin glasses and proteins are still hindered by the multiple-minima problem. It is very difficult for conventional MC and MD methods to sample the relevant parts of configuration space properly at low temperatures. This is because simulations at low temperatures are likely to be trapped in a few out of a huge number of local-minimum-energy states, so that the results are strongly dependent on the initial conditions. One way to surmount this difficulty is to sample configurations upon a non-Boltzmann weighting function instead of the conventional one so that the simulation may easily escape from states of energy local minima.

Several non-Boltzmann sampling methods have been proposed including the umbrella sampling (US) (Valleau and Card, 1972; Torrie and Valleau, 1974, 1977; Valleau, 1993), multicanonical algorithm (MCA) (Berg and Neuhas, 1991; Berg, 1992), multiensemble sampling (MES) (Han, 1992, 1996; Han et al., 2001), entropic sampling (ES),(Lee, 1993) replica-exchange method (REM), (Hukushima and Nemoto, 1996; Hukushima et al., 1996) and so forth. The non-Boltzmann sampling methods are powerful when their weighting functions are chosen properly. But in these methods other than the MES and the REM, the weighting functions are not a priori known and have to be determined numerically by iterations of short trial simulations. In general, this process can be non-trivial and very tedious for complex systems with a number of local-minimum-energy states. In the REM, the weighting function is given as the product of Boltzmann factors, and so it is essentially known. However, this method has a difficulty that other methods do not encounter; as the number of degrees of freedom of the system increases, the required number of replicas also increases markedly, whereas only a single replica is simulated in other methods. This demands a lot of computer power for complex systems. Therefore, there have been suggestions to use the REM only for the determination of weighting functions in other methods (Sugita and Okamoto, 1999; Mitsuak et al., 2001). On the other hand, in the MES, the analytical form of weight function is given as a superposition of normalized Boltzmann factors of the systems. The MES method was originally developed to estimate free energy differences between systems with different potential energies and/or at different thermodynamic states. The MES has not been yet applied to studies of complex molecular systems such as proteins.

Protein folding, i.e. predicting the three-dimensional structure of a protein from its sequence information and elucidating this process, remains an unsolved problem in biophysics and chemistry. It is important because knowledge of the three-dimensional structures and the folding process is essential for understanding the
biological roles and functions of proteins. Recently a conventional MC simulation study of protein folding using a simple united-residue (UNRES) potential model (Liwo et al., 1997) reported that the simulation results depend on the initial conditions of the systems at low temperatures (Kim et al., 2004). Owing to the multiple-minima problem, they failed to obtain the equilibrium ensembles representing the stable phase of native folded states. In this article we demonstrate that the multiple-minima problem in simulations of proteins can be greatly alleviated by implementing the MES to the system. We apply the MES to the equilibrium folding of small proteins using the same potential model UNRES. For each protein, equilibrium thermodynamic properties at several temperatures from the unfolded to the folded states are investigated in a single run of MC simulation at a time. The ability of the MES to explore relevant parts of conformational space is demonstrated in this study.

2 POTENTIAL MODEL AND SIMULATION METHOD

2.1 UNRES energy

In the united-residue (UNRES) model, as shown in Figure 1, a protein chain is represented by a sequence of α-carbon (Cα) atoms connected by virtual bonds with attached united side-chains (SC) and united peptide groups (p) located in the middle of the Cα–Cα virtual bonds. All the virtual bonds are fixed in length; the Cα–Cα length is set to 3.8 Å, and the Cα–SC lengths are given for each amino acid type. The energy of a protein in the UNRES model is given by

\[ E = \sum_{i<j} U_{ai}(i,j) + \sum_{i<j} U_{aj}(i,j) + \sum_{i<j} U_{ap}(i,j) + \sum_{i} \left[ U_{bi}(i) + U_{ci}(i) + U_{di}(i) + \sum_{j \neq i} U_{di}(i,j) \right] \]

where \( U_{ai}(i,j) \) denotes the interaction between the side-chains of residues \( i \) and \( j \), which is expressed by a Lennard–Jones potential. \( U_{aj}(i,j) \) represents the van der Waals interaction between the side-chain of residue \( i \) and the peptide group of residue \( j \), and the potential \( U_{ap}(i,j) \) accounts for the electrostatic interaction between the peptide groups of residues \( i \) and \( j \). The terms \( U_{bi} \), \( U_{ci} \), and \( U_{di} \) represent the short–range interactions corresponding to the energies of virtual angle bending, virtual dihedral angle torsions and side-chain rotamers, respectively. \( U_{di} \) denotes the interaction which forces two cysteine residues to form a disulfide bridge, and the last term \( U_{di}^{(1)} \) represents the four body interaction.

The parameters of the UNRES energy were optimized simultaneously for four proteins (Lee et al., 2004); betanova (20 residues, three-stranded β-sheet) (Bursula and Brooks, 1999; Kortemme et al., 1998), zink-finger based β-β-α motif (28 residues, one β-hairpin and one α-helix) (Dahiyat and Mayo, 1997), HP-36 (36 residues, three-helix bundle) (Duan and Kollman, 1998; Fernander et al., 2003), and fragment B of staphylococcal protein A (46 residues, three-helix bundle) (Zhou and Karplus, 1999; Snejko and Brooks, 2001; Alonso and Daggett, 2000). Using the conformational space annealing method, the parameters were adjusted in such a way that the native-like conformations are more favored than the others energetically (Kim et al., 2003).

2.2 MES

The MES is a powerful algorithm. One can simultaneously investigate several thermodynamic states of systems in a single simulation. The method allows the simulation to explore the relevant parts of configuration space equally for every state of interest. For the simultaneous investigation of

\[ E = \sum_{i<j} U_{ai}(i,j) + \sum_{i<j} U_{aj}(i,j) + \sum_{i<j} U_{ap}(i,j) + \sum_{i} \left[ U_{bi}(i) + U_{ci}(i) + U_{di}(i) + \sum_{j \neq i} U_{di}(i,j) \right] \]

\[ W \propto \left[ \sum_{i=1}^{n} \exp\left(\frac{-\Phi_i}{T_i} - \rho_i \right) \right]^{-\frac{1}{\rho_i}}, \]

\[ C_i = F_i + \text{constant} \]

\[ e^{-\Delta F_i} = \int e^{-\Phi_i} d\Omega = \frac{(W^{-1}e^{-\Phi_i})_W}{(W^{-1}e^{-\Phi_i})_W} = e^{-\Delta C_i} \frac{(f_i)_W}{(f_i)_W}, \]

\[ f_i = \left\{ 1 + \sum_{j \neq i} \exp[-p(\Delta C_j - \Delta \Phi_j)] \right\}^{-\frac{1}{\rho_j}}, \]
where $\Delta \Phi_i = \Phi_i - \Phi_r$. Therefore, starting with an arbitrary set of values for $\Delta C_{ij}$, the iterative replacement of the values using the calculated values of $\Delta F_{ij}$ leads to the self-consistent condition $\Delta C_{ij} = \Delta F_{ij}$. Previously, we showed that a set of near self-consistent values for $C_{ij}$ could be obtained after a few short preliminary runs if the overlaps between configurational distributions of the investigated systems are not so poor (Han, 1992, 1996; Han et al., 2001).

The distributions of conformations relevant to each system can be obtained from the relation

$$\rho_i = \rho_w \frac{f_i}{\langle f \rangle_w},$$  

(6)

where $\rho_w$ is the distribution sampled through the MES. In practice, $\rho_w$ and $\rho_i$ are obtained by accumulating units and values of $f_i$, respectively, into the bins corresponding to the sampled conformations (see subroutine DIST in the Supplementary Materials). The thermodynamic properties of the investigated systems are obtained by calculating canonical ensemble averages of the corresponding quantities from

$$\langle X \rangle_i = \frac{\langle X f \rangle_w}{\langle f \rangle_w}.$$  

(7)

Previously, we examined the dependence of simulation results on $p$ for $p = 0.5, 1, 2$ and 4, and no serious dependence was observed (K.-K. Han and H. S. Son Unpublished data). In this work, we set $p = 1$. Then the weighting function $W$ becomes just the sum of the normalized Boltzmann distributions for the investigated systems;

$$W \propto \rho_1 + \rho_2 + \cdots + \rho_N,$$

(8)

from which one can more clearly see that it is equally probable for the conformations relevant to the individual systems under investigation to be sampled in a MES simulation. Thus, the simulation could easily escape from the states of energy local minima with high energy barriers provided that the systems are taken so that the range of energies of their relevant conformations is wide enough to cover the states of interest and the barriers. There are various ways to achieve this condition; taking systems at a range of thermodynamic states, taking systems with different energy functions and so on. In this work, we took proteins at temperatures from folded states to unfolded states.

3 SIMULATION AND RESULTS

MC simulations using the MES for the folding of betanova and zinkfinger based $\beta$-$\beta$-$\alpha$ motif (hereafter we use its PDB ID, 1fsd) were performed. A parallelized computer code was utilized with 16 processors. Each processor executes the code of MES algorithm independently with its own initial structure of UNRES protein chain. In the simulation, results from the processors are gathered and ensemble averages are calculated. In the UNRES potential model, there are two backbone angles and two side-chain angles per residue; glycines have no side-chains. At each step in our Monte Carlo simulations, one of these angles can have a new value, and the moves involving backbone angles are attempted three times more frequently than those involving side-chains. The maximum ranges of a move are taken as $20^\circ$ and $5^\circ$ for betanova and 1fsd, respectively. The thermodynamic states investigated in a single simulation consist of 11 different temperatures for betanova ($T = 10, 15, 20, 30, 40, 50, 60, 70, 80, 110$ and 150 in arbitrary units) and 13 for 1fsd ($T = 10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 95, 120$ and 150). We performed preliminary runs of the simulation generating $10^6$ configurations were generated for calculations of ensemble averages. We repeated the simulations several times with different initial conformations and random number seeds to verify whether they still give identical results without getting trapped in local minima of energy states.

During the simulations, the root-mean-square deviation (RMSD) from the native structure, the radius of gyration ($R_g$) and the fractions of the native contacts (Q) are measured using $C^\alpha$ coordinates. A native contact is defined to exist when two $C_{\alpha}$s separated by more than two residues in sequence are placed within 7.0 Å. The canonical ensemble averages and distributions of those quantities at each nominal temperature are calculated from Equations (6) and (7). Figure 2 shows the resulting energy distributions. The anomalies of needle-like peaks shown in Figure 2a, which was obtained from a simulation of betanova, indicate that the simulation went through

![](https://academic.oup.com/bioinformatics/article-abstract/22/15/1832/243425/1834)
Application of MES to protein folding

Fig. 3. The ensemble averages of potential energy $U$, RMSD, fractions of native contacts $Q$ and radius of gyration $R_g$ for (a) betanova and (b) 1fsd, as functions of temperature. Curves are plotted in the same frame without specifying the scales on y-axis in order to compare their functional behaviors on the temperature.

states of energy local minima. The simulation does not always go through such states of energy local minima, as shown in Figure 2b obtained from a simulation of 1fsd. It depends on the initial condition of the system. However it appears that the durations of the MES simulation passing through states of energy local minima are so short that it does not affect the results of ensemble averages. Note that all the parts of conformation space relevant to low temperatures are sampled whereas a previous conventional MC simulation study by Kim et al. (2004) failed. The previous study reported that the simulations at low temperatures show non-ergodic glassy behavior ($T/C_20 < 30$ for betanova and $T/C_20 < 50$ for 1fsd). In comparison, we can state that the glassy behavior observed in the previous work is owing to the multiple-minima problem. The calculated ensemble averages as functions of temperature are shown in Figure 3. Their values change drastically on the folding.

In order to understand the thermodynamic behavior of the folding transition more precisely, we also calculated the specific heat, $C_V(T) = \langle (\Delta Q)^2 \rangle /T^2$ and the fluctuations of $Q$, RMSD/$(\text{RMSD})^2$ and $R_g/\text{RMSD}^2$ at individual temperatures $T_i$, i.e. $\langle (\Delta Q)^2 \rangle$, $\langle (\Delta \text{RMSD})^2 \rangle /\text{RMSD}^2$ and $\langle (\Delta R_g)^2 \rangle /\text{RMSD}^2$. Before completing the calculations, we incorrectly guessed, without a theoretical basis, that the temperature dependence of $\langle (\Delta Q)^2 \rangle$ would be similar to the specific heat while $\langle (\Delta Q)^2 \rangle /\text{RMSD}^2$ would be similar to that of $\langle (\Delta \text{RMSD})^2 \rangle /\text{RMSD}^2$. The results are plotted in Figure 4 as functions of temperature. The plots show that the folding processes of both proteins consist of two consecutive structural transitions; a collapse transition from random coil conformations to compact conformations represented by the peaks of specific heat and the fluctuation in the fraction of native contacts $Q$, and a folding transition from compact conformations to native

1835

Fig. 4. The heat capacity $C_V = \langle (\Delta Q)^2 \rangle$, $D_{\Delta Q} = \langle (\Delta Q)^2 \rangle /\text{RMSD}^2$ and $D_{\Delta R_g} = \langle (\Delta R_g)^2 \rangle /\text{RMSD}^2$ for (a) betanova and (b) 1fsd, as functions of temperature. Curves are plotted in the same frame without specifying the scales on y-axis in order to compare their functional behaviors on the temperature.
increases as the temperature decreases. the collapsed phase and the frequency of native conformations non-native compact confirmations and native-like conformations in and after the transition. These distributions show that there coexist illustrates the distributions of RMSD at the temperatures before transition from the collapsed phase to the native phase. Figure 5 RMSD and used as a structural overlap function in this work. The results for (Camacho and Thirumalai, 1993; Socci and Onuchic, 1995; Guo and Brook, 1997) RMSD/h. This feature of two consecutive transitions has been observed in minimalist model simulations (Camacho and Thirumalai, 1993; Socci and Onuchic, 1995; Guo and Brook, 1997; Lockner and Hernander, 2004). As in the previous works using minimalist model where a structural overlap function was defined to measure the degree of disorder of the protein with respect to the native state and the peak of its fluctuation was used to identify the folding transition, (Camacho and Thirumalai, 1993; Socci and Onuchic, 1995; Guo and Brook, 1997) RMSD/RMSD is used as a structural overlap function in this work. The results for RMSD and Rg shown in Figures 3 and 4 clearly indicate a structural transition from the collapsed phase to the native phase. Figure 5 illustrates the distributions of RMSD at the temperatures before and after the transition. These distributions show that there coexist non-native compact confirmations and native-like conformations in the collapsed phase and the frequency of native conformations increases as the temperature decreases.

4 SUMMARY

We applied the MES algorithm to the equilibrium folding of betanova and 1fsd. The proteins at several temperatures from the unfolded to the folded states were simulated in a single MC run at a time and their equilibrium thermodynamic properties were calculated correctly. The distributions of conformations sampled in our MES MC simulations clearly indicate that, when going through states of energy local minima, the MES simulation did not get trapped in them but escaped from them so quickly that all the relevant parts of conformation space could be sampled properly. The calculational results show that, for both proteins, folding takes place via two consecutive structural transitions consisting of a collapse transition followed by a folding transition. From these results, we can conclude that the use of MES in simulations of complex molecular systems like proteins greatly alleviates the multiplet-minima problem and that the MES is especially useful in studying the equilibrium folding of proteins.

ACKNOWLEDGEMENTS

This work was supported by the Korea Research Foundation Grant(KRF-2004-C00044), in which main Calculations were performed by using the supercomputing resources of the Korea Institute of Science and Technology Information (KISTI).

Conflict of Interest: none declared.

REFERENCES
