Journal of Natural Sciences
Pai Chai University, Korea
Vol. 18, No. 1 : 1-8, 2007

THE Multiensemble Sampling Method

Kyu-Kwang Han
Department of Physics, Pai Chai University, Daejeon 302-735, Korea

다중앙상블 표본추출 방법

한 규 광
배재대학교 전산전자물리학과, (우) 302-735 대전시 서구 도마동 439-6

ABSTRACT

An efficient sampling method of computer simulation is reviewed. Using the method, several thermodynamic states can be investigated at a time in a single simulation. It is due to the ability of the method to explore the relevant parts of configuration space equally for every state being investigated. The method can be used in simulations of complex systems such as biopolymers which are still greatly hampered by the multi-minima problem. In this article I present a brief theoretical review of the method and illustrate how to realize it in the simulations.

요 약

컴퓨터 시뮬레이션의 효율적인 한 방법을 재조명하였다. 이 방법을 이용하면, 여러 열역학적 상태들을 단 한 번의 시뮬레이션으로 조사할 수 있다. 그렇게 할 수 있는 것은, 조사하고자 하는 모든 상태들에 대해 관련 배열공간을 골고루 탐사하는 방법의 능력에 기인한다. 이 방법은 아직도 다중최소 문제가 여전히 큰 장애로 남아 있는 생체고분자와 같은 복잡계의 시뮬레이션에도 이용할 수 있다. 이 논문에서 방법의 이론을 간단히 재검토하고 어떻게 시뮬레이션으로 실현하는지 예를 들어 설명하겠다.

Key words: computer simulation, multiminima problem, free energy, protein folding

---

Corresponding Author : Kyu-Kwang Han, Department of Physics, Pai Chai University, Daejeon Korea, 302-735, Tel : +82-42-520-5464, E-mail : khan@pcu.ac.kr

1. INTRODUCTION

The computer experiment employing Monte Carlo (MC) or molecular dynamics (MD) is now a popular tool in science. Despite the recent development of supercomputing power, simulations of larger complicated systems such as biopolymers are still hindered by the multi-minima problem. It is very hard for conventional MC and MD methods to sample the relevant configurations properly at low temperatures. This is because simulations at low temperatures would likely be trapped into a few of a huge number of local-minimum-energy states so that the results will strongly depend on the initial configuration. One way to surmount the difficulty is to sample configurations upon a non-Boltzmann weighting function instead of the conventional Boltzmann weighting function so that the simulation may escape from trapped local minima.

There have been proposed several non-Boltzmann sampling method; the umbrella sampling method (USM) [1-4], multicanonical method [5,6], multiensemble sampling method (MESM) [7-9], entropic sampling method (ES) [10,11] replica-exchange method (REM) [12,13], etc. The non-Boltzmann samplings method are powerful when their weighting functions are properly chosen, but for the above methods except the MESM and the REM, they are not a priori known and have to be determined by iteration of preliminary simulations. This process is, in general, nontrivial and very tedious for complex systems.

In the REM, on the contrary, the weighting function is given as the product of Boltzmann factors. A number of non-interacting replicas of a system at different thermodynamic states are simulated independently and simultaneously by the conventional MC or MD. Every few steps, pairs of replicas are exchanged with a specified transition probability. However, the method has a difficulty that the other methods do not encounter as the number of degrees of freedom of the system increases, the required number of replicas also greatly increases, whereas only a single replica is simulated in the other methods. This demands a lot of computer power for complex systems.

In the MESM, on the other hand, the weight function is given as a superposition of Boltzmann factors and only a single replica is simulated. Originally, as like as the USM, the method was developed for the accurate estimation of the free energy and has not been applied

yet to studies of complex molecular systems such as proteins. Recently, we have demonstrated the feasibility of extending the method to simulate the protein folding. [14] In this paper, we present a brief review the method and illustrate how to realize it in the applications to the hydration free energy and the protein folding.

2. THEORY

For the sake of a logical argument, we begin by reviewing the basic theory of non-Boltzmann sampling scheme. Suppose that there are n similar systems with potential energies U_l at temperatures T_l, $l = 1,\cdots,n$, and one is going to investigate them in a single simulation. It is impossible, in general, to obtain data of other n-1 system from the conventional MC or MD simulation for one of the systems since the parts of configuration space sampled upon the Boltzmann weight in the simulation are not broad enough to cover all the parts of configuration space relevant to the other systems.

For the successful simultaneous investigation, configurations should be sampled upon a general non-Boltzmann weighting function W which covers all the parts of configuration space relevant to the investigated systems. The probability density of configurations, ρ_l, for system l is related with the probability density of sampled configurations, ρ_w, by

$$\begin{aligned}\rho_l &= \frac{e^{-\Phi_l}}{\int e^{-\Phi_l} d\Omega} \\ &= \frac{W^{-1}e^{-\Phi_l}W / \int W d\Omega}{\int W^{-1}e^{-\Phi_l}W / \int W d\Omega} \\ &= \frac{W^{-1}e^{-\Phi_l}}{\langle W^{-1}e^{-\Phi_l}\rangle_W}\rho_W \end{aligned} \tag{1}$$

where $\Phi_l = U_l / kT_l$ with k being the Boltzmann constant, $d\Omega$ represents the volume element in the configuration space and $\langle\ \rangle_W$ denots an average over sampled configurations. The canonical ensemble average of a physical quantity X for system l is calculated by

$$\langle X \rangle_l = \int X \rho_l d\Omega = \frac{\langle X W^{-1} e^{-\Phi_l} \rangle_W}{\langle W^{-1} e^{-\Phi_l} \rangle_W} \tag{2}$$

With an appropriate choice of W, one would be able to investigate several systems in a single simulation. But, the matter is how to choose W appropriately. The efficiency of a non-Boltzmann sampling simulation is determined by the choice of W. The job without an a priori recipe for W is uncertain.

In the MESM, the weighting function W is given by

$$W = \left[\sum_{l=1}^{n} e^{p(C_l - \Phi_l)} \right]^{1/p} \tag{3}$$

In Eq. (3), p is an arbitrary constant which does not affect the calculation results. The dependence of simulation results on p has been checked by setting p = 0.5, 1, 2, 4 in our previous work and no serious dependence was observed. [14] On the contrary, the adjustable parameters C_l critically determine the distribution of sampled configurations. For example, if one took $C_l >> (<<) C_l$ for every $i \neq l$ and , the relevant configurations for the system l only will (won't) be sampled.

Originally, Eq. (3) with p = 2 was derived by the functional minimization of the sum of the squares of expected relative errors in the denominator in the most right hand side of Eq. (1) so that the canonical distribution ρ_l can be calculated with equal accuracy for every system.[7,8] As shown in the original derivation, it is optimal for the simulation when $C_l = F_l + const$ where F_l is the temperature scaled free energy of system l. Note that Eq. (3) becomes a superposition of the normalized canonical distributions of configurations relevant to individual systems then;

$$W = const \times \left[\sum_{l=1}^{n} \rho_l{}^{p} \right]^{1/p} \tag{4}$$

Note that W is just the sum of normalized canonical distributions of configurations relevant to individual investigated systems when one takes p = 1.

3. REALIZATION

In order for the method to work, one needs to know the values of the free energies of investigated systems. They can be obtained by calculating

$$e^{-\Delta F_{lm}} = \frac{\int e^{-\Phi l} d\Omega}{\int e^{-\Phi m} d\Omega} = \frac{\langle W^{-1} e^{-\Phi l} \rangle_W}{\langle W^{-1} e^{-\Phi m} \rangle_W} \tag{5}$$

where $\Delta F_{lm} = F_l - F_m$. Thus, starting with an arbitrary set of values for C_l, the iterative replacement of the value by the calculated of ΔF_{lm} in a simulation leads to the self-consistent condition of $C_l - C_m = \Delta C_{lm} = \Delta F_{lm}$. Let us illustrate this by taking an example of sampling two ensembles in a single simulation as shown in Fig. 1. The two systems are an uncharged spherical particlein water (system 0) and a partially charged one by 0.3*e* in water (system 1). The free energy difference of the systems is -12, but it is not *a priori* known. Preliminary runs for the adjustment of ΔC_{10} were done, started with $\Delta C_{10} = 0$ the value of ΔC_{10} was replaced iteratively by the estimate of ΔF_{10} obtained in the preceding run. In each run, 10^7 configurations were generated. The estimates of ΔF_{10} after the second run oscillate about its true value; less than 5% deviation. In Fig. 1 are plotted the distributions of configurations sampled in the preliminary runs. They confirm the theory that two ensembles are equally sampled when the self-consistent condition is satisfied.

The method has been applied to the free energy of charging a sodium ion in water. [9] Eleven states, whose charges on the ion are from 0 to 1*e* in 10 steps of 0.1*e*, are included for calculating the free energy. A set of near self-consistent values for C_l is obtained by routines of two-ensemble sampling runs for the pairs of nearest states first, in each of which 10^6 configurations were generated. Using the set of these values, 3 x 10^7 configurations were generated by a run of sampling 11 ensembles to calculate free energies. An additional run of generating 5 x 10^7 configurations was done to confirm that the self-consistent condition is satisfied. Figure 2 shows the results.

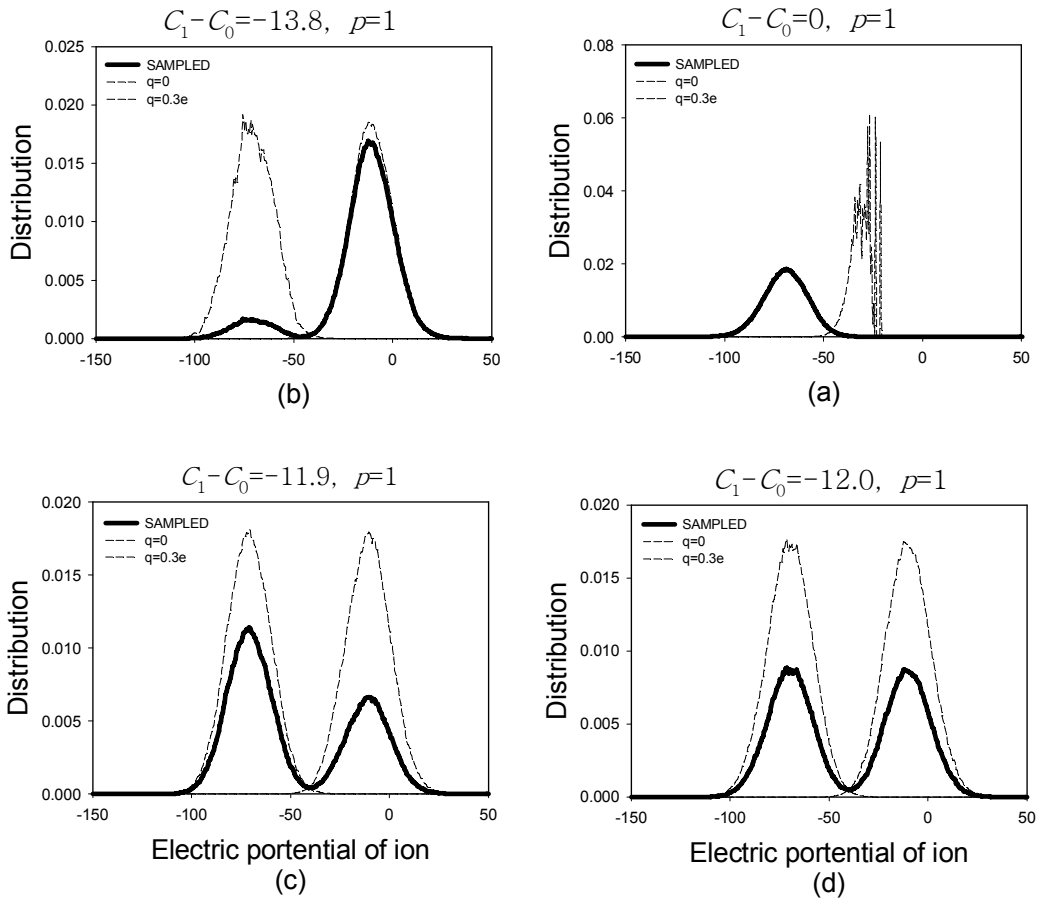

FIGURE 1. Distributions of the ionic electric potential due to water molecules obtained in the preliminary runs of sampling two ensembles. As the preliminary run with replacement of ΔC_{10} by ΔF_{10} is repeated, the parts configurations relevant to individual systems are equally sampled; (a) → (b) → (c) → (d). The thick solid curves are the sampled distributions and the thin dashed curves are the normalized canonical distribution calculated in simulations.

Recently we also applied the method to the folding of small proteins.[14] A simple united-residue (UNRES) [15] potential model was used in the simulations. In the work we started with simulations of sampling the ensembles of two high temperatures at which the protein remains unfolded. We increased the number of ensembles being sampled one by one lowering thetemperature to the folded phase. Figure 3 where the resulting distributions of

energy and RMSD(root-mean-square deviation) for betanova are plotted shows that the ensembles of the temperatures being invetigated are sampled equally.

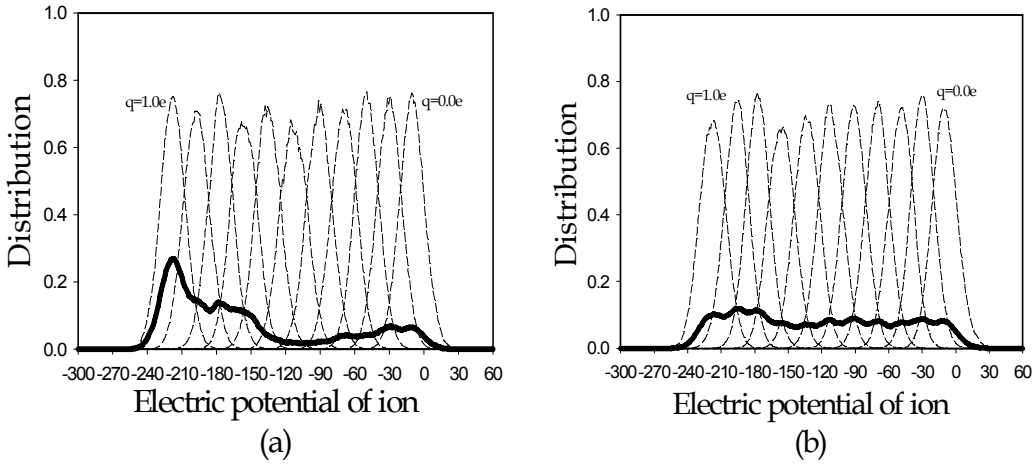

FIGURE 2. Distributions of the ionic electric potential due to water molecules obtained in the runs of sampling eleven ensembles; (a) the distribution sampled in the simulations using a set of values for C_l that is obtained by routines of two-ensemble sampling runs for pairs of nearest systems, (b) the distribution sampled in the simulations with replacement of C_l by F_l that are calculated in the simulation (a).

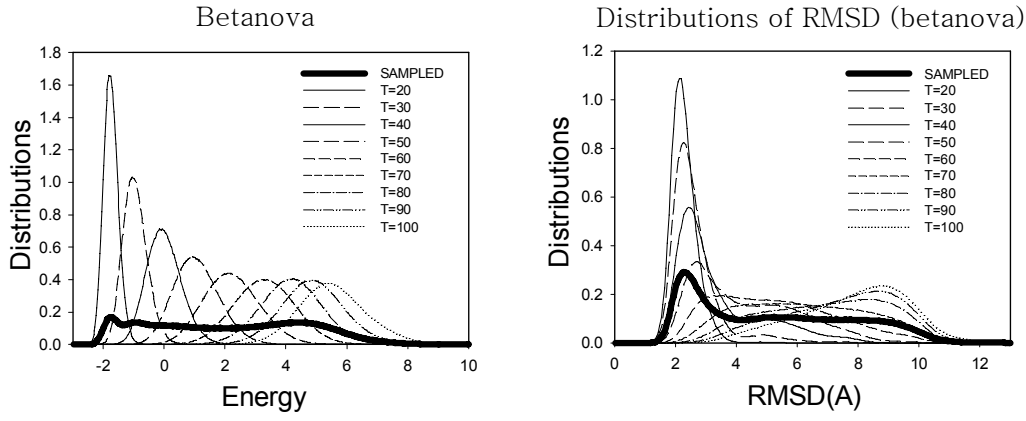

FIGURE 3. Distributions of energy and RMSD obtained in the simulations of the folding of betanova using the MESM.

4. SUMMARY

The multiensemble sampling method has been reviewed briefly with demonstrating the method is able to explore all the parts of configuration space relevant to the systems of interest. In the method, the weighting functione W is a superposition of Boltzmann factors of the systems. It is simple and easy a set of optimal values for the parameters of W. The method can be used in simulations of complex systems which are still greatly hampered by the multi-minima problem. We are extending the method to simulate the folding of proteins.

REFERENCES

1. J. P. Valleau and J. Card, *J. Chem. Phys.* 57(1972), 5457.
2. G. Torrie and J. P. Valleau, *Chem. Phys. Lett.* 28(1974), 578.
3. G. Torrie and J. P. Valleau, *J. Comp. Phys.* 23(1977), 187.
4. J. P. Valleau, *J. Chem. Phys.* 99(1993), 4718.
5. B. A. Berg and Neuhaus, T. Phys. Lett. B267(1991), 249-253.
6. B. A. Berg, Phys. Rev. Lett. 68(1992), 9-12.
7. K.-K. Han, *Phys. Lett. A*, 165(1992), 28.
8. K.-K. Han, *Phys. Rev. E*, 1996, 54, 6906.
9. K.-K. Han, K. H. Kim, B. J. Mhin, and H. S. Son, *J. Compu. Chem.* 22(2001), 1004.
10. J. Lee, *Phys. Rev. Lett.* 71(1993), 211-214.
11. J. Lee, *Phys Rev Lett* 71(1993), 2353.
12. K. Hukushima and K. Nemoto, *J. Phys. Soc. Jpn.* 65(1996), 1604-7608.
13. K. Hukushima, H. Takayama, and K. Nemoto, *Int J Mod Phys* C 7(1996), 337-344
14. S. H. Son, S.-Y. Kim, J. Lee and K.-K. Han, *Bioinformatics* 22(2006), 1832-1837
15. A. Liwo, S. Oldziej, M. R. Princus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga, *J. Comp. Chem.* 18(1997), 849