Journal of Molecular Graphics and Modelling 27 (2008) 321–325

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

A simple method of estimating sampling consistency based on free energy map distance

Won-Joon Son [a], Soonmin Jang [b], Seokmin Shin [a,*]

[a] School of Chemistry, Seoul National University Seoul 151-747,Republic of Korea
[b] Department of Chemistry, Sejong University, Seoul 143-747, Republic of Korea

ARTICLE INFO

Article history:
Received 5 January 2008
Received in revised form 20 May 2008
Accepted 23 May 2008
Available online 10 July 2008

Keywords:
Replica exchange method
Free energy
Sampling consistency

ABSTRACT

Free energy surfaces, calculated during computer simulations, are known to be useful in characterizing the system of interest such as bio-molecules. However, it is usually very difficult to evaluate free energy from direct simulations, mainly because of high computational costs. Several simulation strategies, including replica exchange method (REM), have been developed to overcome this problem by providing efficient conformational sampling methods. Even with such efficient simulation schemes, fundamental questions concerning simulation convergence still remain to be resolved. In this paper, we propose to use a meta-distance between different free energy surfaces as one of the minimal measures for determining simulation consistency. This method is used for examining free energy surfaces obtained from folding simulations of a synthetic 11-residue protein (1AQG) using REM.

1. Introduction

Conformation sampling method based on the replica exchanges, called as REM, is one of the widely adopted techniques for simulating many complex molecular systems, including proteins folding, to overcome notorious trapping problems. Convergence properties of trajectories for various sampling methods are usually very difficult to be confirmed. An identification of the global convergence of simulations has never been a trivial problem, and will be a more challenging problem as much longer simulations on complex systems are attempted. A measure based on the fluctuation of the cosine content of the principal components would be one of the candidates [1], which is, however, rather cumbersome to be applied to replica exchange trajectories. In general, the convergence identification is more complicated for REM, since coordinates and velocities are exchanged between different replicas during the simulation.

Besides its powerful sampling efficiency, one of the reasons behind popular use of REM is that it can generate canonical properties unlike other *ad hoc* simulation strategies based on non-general ensemble schemes. Therefore, the convergence of the simulation has crucial importance to obtain reasonable canonical ensemble properties. Despite of its importance, the convergence issue in REM is rather not well studied even though there have been several attempts to improve REM efficiency [2,3]. Usually, REM simulations spanning from couple of nano-seconds to about 100 ns are used to calculate many canonical properties such as heat capacity, average energy, and free energy, after discarding some initial trajectories. However, this process could give somewhat poor results, without proper checking of the simulation convergence. There could be many different criteria, which show that the simulation has reached equilibrium, *i.e.* converged. Strictly speaking, only the infinitely long time simulation can tell if the simulation has reached equilibrium in true sense [4,5]. In their GB1 folding study, Brooks et al. [6] checked the simulation convergence by performing REM simulations starting from both folded and unfolded conformations and comparing the results from the two simulations. In principle, the equilibrium (canonical) properties should be the same regardless of initial starting conformations if the simulations are converged. It is known that the free energy barrier crossings, constructed from simulations, take rather long time even with the REM simulation [2,7]. In this aspect, free energy convergence can reflect the quality of REM simulation and any quantitative measure of its convergence can greatly assist the validation of simulations.

If the convergence problem is restricted to the trajectories sampled and to the target properties, analysis of the convergence of representative features of the sampled data set is much more manageable. Restricted convergence in this sense may be called as "sampling consistency". In protein folding problems, pseudo-free energy surface, usually denoted as a contour map of two degrees of freedom, is one of the most compact and efficient way of

* Corresponding author. Fax: +82 2 889 1568.
E-mail address: sshin@snu.ac.kr (S. Shin).

1093-3263/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2008.05.006

visualizing global features of the whole simulation data. The shape of the free energy map is often used for providing an evidence for internal convergence of REM-based trajectories. Any quantitative measure, which indicates a similarity between the two free energy maps, will make analysis more plausible. In this paper, we propose to use a meta-distance between two free energy maps as a quantitative measure for the similarity.

2. Methods

Consider the two different free energy surfaces. Any measure which provides quantitative value concerning the difference between the two free energy surfaces can be utilized. We adopted the quantity, $\tilde{\Theta}$, which is based on complementary quantity of cosine similarity, as the meta-distance between free energy map X and Y:

$$\tilde{\Theta} = 1 - \cos(\vec{X}, \vec{Y}) = 1 - \frac{\vec{X} \cdot \vec{Y}}{|\vec{X}| \cdot |\vec{Y}|} \quad (1)$$

Cosine distance $\cos(\vec{X}, \vec{Y})$ is one of the simplest distances in pattern recognition [8]. The complementary cosine distance defined above, which gives 0 for the two identical maps, is more intuitive in comparing the similarity among free energy surfaces. Here vectors $\vec{X}$ and $\vec{Y}$ are isomorphic to the free energy maps X and Y, respectively. $\vec{X}$ and $\vec{Y}$ can contain either simply the frequencies ordered with respect to the grid index or the free energy values, and corresponding maps X and Y are essentially matrices representing free energy surfaces. The dimension of the vector is identical with the total number of grid cells in the free energy map. Because the free energy map is constructed from histogram, we already have frequency data on each grid cell. In practice, grids outside the reference histogram may be neglected. It is noted that the complementary cosine distance is unitless and various physical quantities can constitute $\vec{X}$ or $\vec{Y}$, even with additional weight factors. It is difficult to give quantitative interpretations for such meta-distance in terms of physical quantities such as (free) energy. In the case of the complementary cosine distance for the free energy surfaces, the distance may suggest the fraction of total free energy difference between the two surfaces.

We applied this measure to a REM trajectory of 1AQG, a synthetic 11-residue protein segment (IKENLKDCGLF) derived from the heterotrimeric GTP-binding protein transducin α subunit C-terminal GTα (340–350) [9]. Traditional Replica-Exchange Molecular Dynamics simulation was performed with the molecular dynamics package TINKER [10]. The force field we used was CHARMM27 with Generalized Born solvent-accessible surface (GBSA) by Case et al. [11] as implemented in TINKER. Velocity version of Verlet integrator with RATTLE was used. Time step was 2 fs. It is found that eight replicas (300.0 K, 327.0 K, 358.0 K, 392.0 K, 431.0 K, 473.0 K, 520.0 K, and 578.0 K) are sufficient for the successful folding simulation. The simulation time length of each replica was 48 ns (a total of 384 ns with all the replicas) and the trajectory was saved at every 200 ps for further analysis.

The free energy surface was constructed from the resulting population distribution using

$$\Delta F(\mathrm{RMSD}, R_g) = -RT \ln P(\mathrm{RMSD}, R_g) \quad (2)$$

where $\Delta F(\mathrm{RMSD}, R_g)$ is the free energy difference for the state with radius of gyration R_g and root mean square deviation (RMSD) based on all the atoms except hydrogen. Here, RMSD was calculated with respect to the PDB structure of 1AQG and R_g is the atomic mass weighted radius of gyration. R is the gas constant, T is the temperature, and $P(\mathrm{RMSD}, R_g)$ is the normalized population. Conformations from the initial simulation data are included for free energy construction. In general, proteins in physiological conditions are of main concern and only the trajectories corresponding to the temperature of 300 K will be used for the analysis. Since the trajectories are exchanged among different temperatures during the REM simulation, the sampling consistency of the replica for 300 K can be considered as a representative for the whole trajectories at higher temperatures.

3. Results and discussions

With eight replicas, we performed conformation searching up to 48 ns on each replica from the fully extended (and minimized) linear configuration, which corresponds to total 384 ns simulation time. It is a common practice to check the convergence of the first stage of simulations by examining quantities which evolve in time with short enough correlation time, such as root mean squared distance or radius of gyration. But the representative quantities may have different time scales, and often we need quantities which require information for the whole trajectory considered. For example, the principal component analysis (PCA) usually requires whole trajectory data within a certain interval. The behavior of principal components, obtained from parts of trajectories, may be used to examine the convergence of the simulations. Kubitzki and de Groot showed that the time evolution of the sampled configuration space area, measured using projections of the trajectory into the first two major PC, can serve as a measure of convergence for REM simulations [12]. In the present study, we examined the behavior of the simplest geometric quantities, atomic mass weighted radius of gyration R_g and the root mean squared distance (RMSD). Meta-distance measure on free energy surfaces, reflecting the time evolutions of such quantities, will be used as a preliminary convergence test. All the following analysis is applied only to the trajectories at 300 K. Fig. 1 shows the time evolutions of R_g and RMSD. Free energy surface as a function of R_g and RMSD, obtained from the whole 48 ns trajectories is shown in Fig. 2.

We build up a free energy surface of (RMSD, R_g) with varying time interval, or "window." We shifted the window by half the size of the window, covering the whole simulation time. Reference free energy map is set as that of the last window data. Fig. 3 shows the complementary cosine distance ($\tilde{\Theta}$) between the reference free energy map and those of shifting windows for each window size. The frequencies ordered with respect to the grid index are taken as the components. It is noted

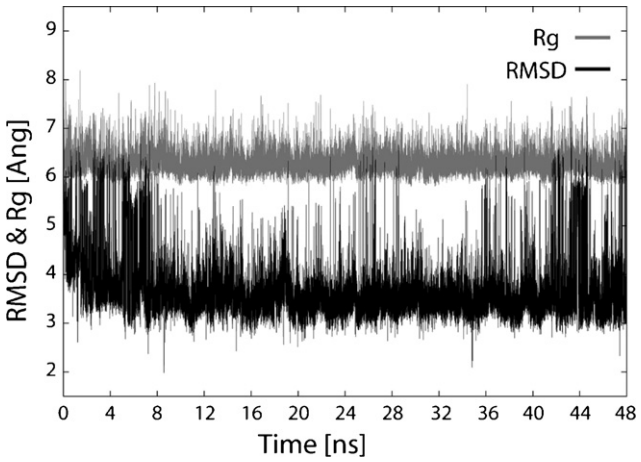

Fig. 1. Time evolutions of R_g and RMSD for trajectories of REM simulations at 300 K. Data at 0 ns corresponds to the fully extended initial structure.

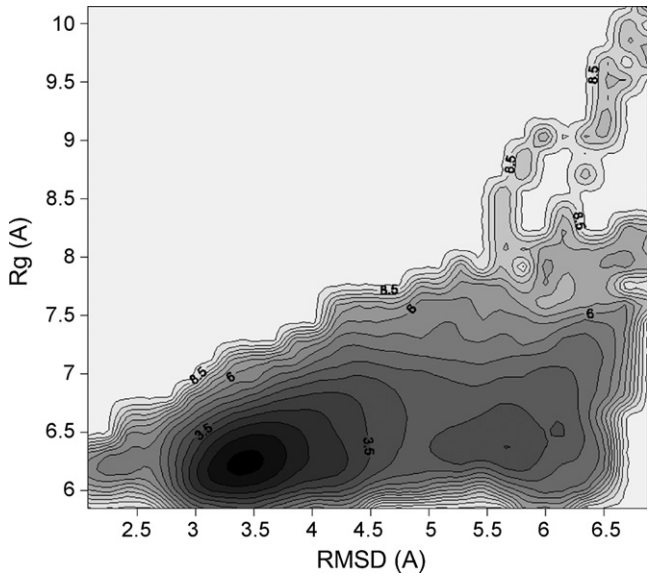

Fig. 2. Free energy surface as a function of R_g and RMSD, obtained from the whole 48 ns trajectory of the REM simulation at 300 K.

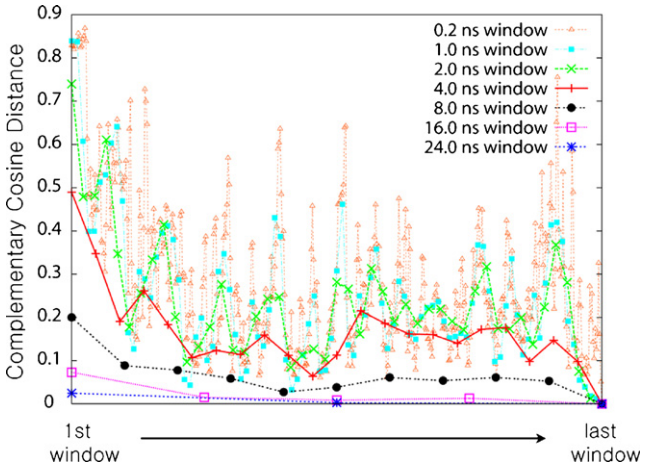

Fig. 3. Complementary cosine distances ($\tilde{\Theta}$) between the reference free energy map and those of the shifting windows for varying window size. The last window in each window size is taken as the reference.

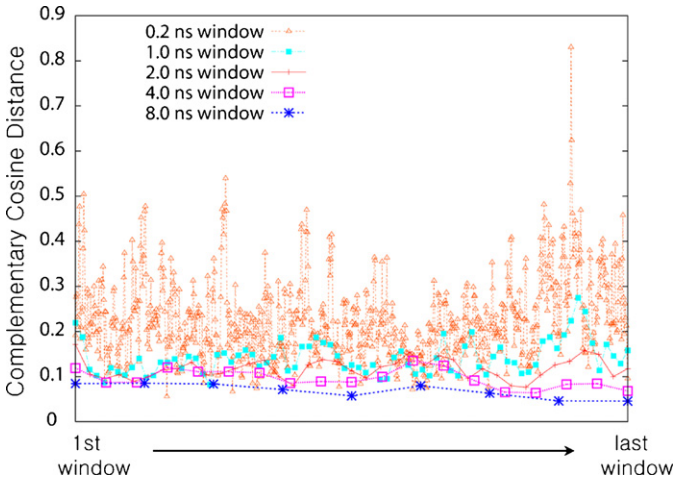

Fig. 5. Complementary cosine distances ($\tilde{\Theta}$) between the reference free energy map and those of the shifting windows for varying window size. The reference is the free energy surface constructed with the whole trajectories from 8.0 ns up to 48.0 ns.

that the distance is decreasing as the window size increases. Distances for 2.0 ns window still show somewhat oscillatory feature. The oscillations are much more reduced for 4.0 ns window and global fluctuations are mostly smoothed out beyond the 8.0 ns window size. Fig. 4 shows the evolution of free energy surfaces as a function of (RMSD, R_g) for (a) 2.0 ns, (b) 8.0 ns, (c) 16.0 ns and (d) 24.0 ns windows. As the window size increases, overall shapes of maps become converged. Look also the similarity between the free energy surface in Fig. 2 and those from the final windows in Fig. 4. It is noted that the reference window, taken as the last window, has no information on the initial collapse from the extended chain. Since the data outside the region of the reference window are not included, the upper right region of Fig. 2, corresponding to the extended configuration, is not reproduced in Fig. 4. Fig. 5 shows the complementary cosine .distance with respect to the free energy surface constructed with the whole trajectories from 8.0 ns up to 48.0 ns. Each of the first sliding windows covers trajectories starting from 8.0 ns. In this case, free energy values, instead of the frequencies, are used. The early trajectories up to 8.0 ns are excluded because the low frequency regions come to play dominant role in distance determination when the trajectories corresponding to the upper right region of Fig. 2 are included. The complementary cosine distances in this case are relatively smaller than those of Fig. 3, suggesting that the free energy of

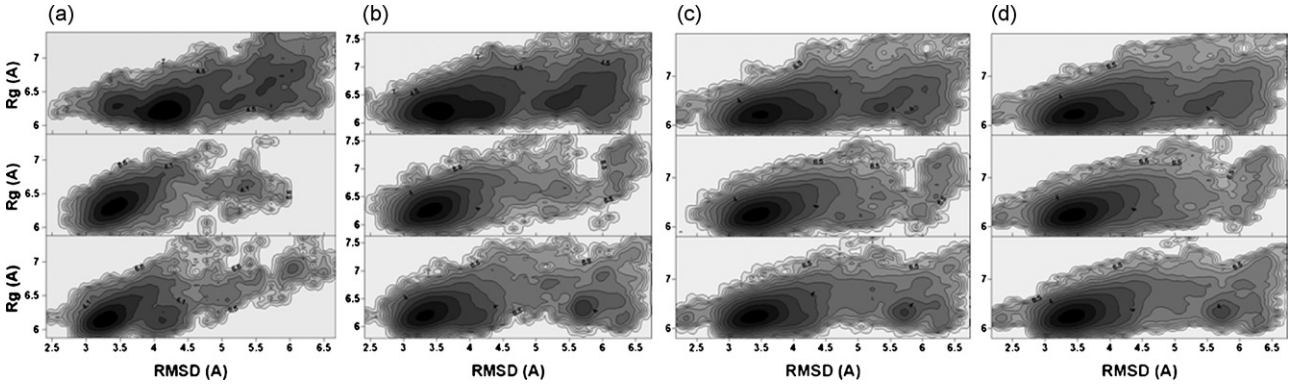

Fig. 4. Free energy surfaces as a function of (RMSD, R_g) for (a) 2.0 ns, (b) 8.0 ns, (c) 16.0 ns and (d) 24.0 ns window sizes. Top, middle, and bottom contours represent the free energy surfaces corresponding to the first, the midst, and the last windows, respectively. Complementary cosine distances with respect to the corresponding last window (bottom contour) are (a) (0.740, 0.282), (b) (0.200, 0.038), (c) (0.073, 0.008), (d) (0.025, 0.003) for (top, middle) contours of each panel.

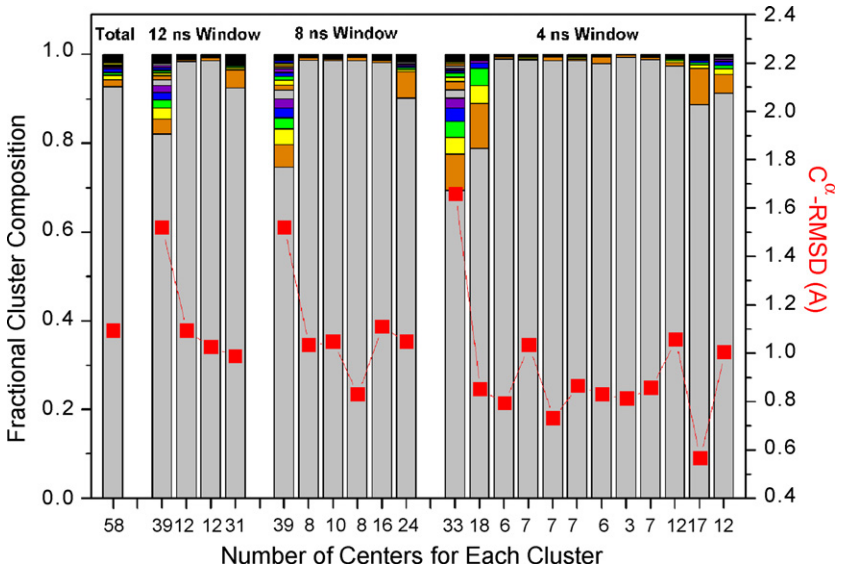

Fig. 6. Cluster analysis for conformations represented by different windows for the window sizes of 48 ns (whole trajectory), 12 ns, 8 ns, and 4 ns, respectively. The number of cluster centers for each window is also given. Red square designates C^{α}-RMSD value for the structure of the center of the largest cluster with respect to the free energy minimum structure.

the last window may not be a good reference. From the distance criteria shown in Figs. 3 and 5, it may be concluded that 8.0 ns is a minimal window size to provide representative features of the present simulation data.

In order to examine the characteristic structures represented by different window sizes, we performed a cluster analysis (Fig. 6). Conformations from every 15th frame of the whole 48 ns trajectories were chosen for the cluster analysis. For cluster analysis, we allow no overlap between neighboring windows. Each cluster center was identified with criteria of C^{α}-RMSD (1.75 Å), and the number of centers for clustering of each window is also denoted in Fig. 6. Except the early stage of the simulation, the single largest cluster dominates the conformations. We found slight increases in the number of cluster centers, even though contributions of minor clusters remained small. One may argue that our sample trajectory has not been fully converged after 48 ns. We calculated RMSD values between the structure of the center of the largest cluster and that of the free energy minimum at (RMSD, R_g) = (3.428, 6.252) in Fig. 2. The fluctuations of the RMSD values are relatively small for large window sizes. The converged values of the RMSDs are similar to the RMSD of the largest cluster for the whole (48 ns) trajectory. The small difference is again attributed to the early collapse stage of fully extended chain which is included only for the analysis of the full trajectory. The results of the cluster analysis are generally consistent with those from the simple convergence tests based on meta-distance approach.

In this paper, we presented a simple method of estimating the similarity of free energy surfaces for different time intervals, thereby checking the sampling consistency of the corresponding simulations. The free energy constructed from histogram has uncertainty depending on the number of grids used. It is emphasized that the convergence of simulations can be assessed only after sufficient simulation steps. It may be possible that the simulation may progress slowly, with respect to certain collective variables of interest, for some time intervals due to the nature of dynamics of the system such as being trapped in local minima. During such stage of trajectory, the distance matrices show small value, suggesting the convergence of the simulation. Only further simulations (either by extending to longer time simulation or performing accelerated simulation) can tell the true convergence of the system. Therefore, the quantitative free energy distance estimation introduced here might not be used as absolute criteria for the simulation convergence. Nevertheless, we believe the meta-distance approach will make not only a rule-of-thumb guide to the internal convergence of simulation, but also a quantitative standard to the comparison of the efficiency of different methodologies. The distance between the two multi-dimensional free energy surface, such as F(RMSD, R_g, number of native contact, number of hydrogen bonding, . . .), also can be measured using the present approach. We also should note the recent work done by Lyman and Zuckerman [13]. Their analysis is based on the structural diversity, not on the free energy as we have done here. It can be argued that the meta-distance approach is robust enough to be applied to the structural diversity-based histogram.

Acknowledgments

This work was supported by grant R01-2006-000-10418-0 from the Basic Research Program of the Korea Science & Engineering Foundation. This work was also supported by a grant from Marine Biotechnology Program funded by Ministry of Land, Transport and Maritime Affairs, Republic of Korea.

References

[1] B. Hess, Convergence of sampling in protein simulations, Phys. Rev. E (2002) 031910.
[2] A. Barducci, R. Chelli, P. Procacci, V. Schettino, F.L. Gervasio, M. Parrinello, Metadynamics simulation of prion protein: structure stability and the early stages of misfolding, J. Am. Chem. Soc. 128 (2006) 2705–2710.
[3] H. Li, G. Li, B.A. Berg, W. Yang, Finite reservoir replica exchange to enhance canonical sampling in rugged energy surfaces, J. Chem. Phys. 125 (2006) 144902.
[4] A. Leach, Molecular Modelling: Principles and Applications, Prentice Hall, 2001.
[5] T. Schlick, Molecular Modeling and Simulation, Springer, 2002.
[6] J. Chen, W. Im, C.L. Brooks III, Balancing solvation and intramolecular interactions: toward a consistent generalized Born force field, J. Am. Chem. Soc. 128 (2006) 3728–3736.
[7] R. Zhou, Trp-cage: folding free energy landscape in explicit water, Proc. Natl. Acad. Sci. U.S.A. 100 (2003) 13280–13285.

[8] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, Wiley-Interscience, 2000 .
[9] O.G. Kisselev, J. Kao, J.W. Ponder, Y.C. Fann, N. Gautam, G.R. Marshall, Light-activated rhodopsin induces structural binding motif in G protein α-subunit, Proc. Natl. Acad. Sci. U.S.A. 95 (1998) 4270–4275.
[10] J. W. Ponder, TINKER—Software Tools for Molecular Design. http://dasher.wustl.edu/.
[11] A. Onufriev, D. Bashford, D.A. Case, Exploring protein native states and large-scale conformational changes with a modified generalized Born model, Proteins: Struct. Funct. Bioinf. 55 (2004) 383–394.
[12] M.B. Kubitzki, B.L. de Groot, Molecular dynamics simulations using temperature-induced essential dynamics replica exchange, Biophys. J. 92 (2007) 4262–4270.
[13] E. Lyman, D.M. Zuckerman, Ensemble-based convergence analysis of biomolecular trajectories, Biophys. J. 91 (2006) 164–172.