John von Neumann Institute for Computing

Modeling of Water Explicitly in the Replica-Exchange Simulation Method for Protein Folding

Mu Yuguang, Yang Ye

published in

NIC Workshop 2006,
From Computational Biophysics to Systems Biology,
Jan Meinke, Olav Zimmermann,
Sandipan Mohanty, Ulrich H.E. Hansmann (Editors)
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **34**, ISBN-10: 3-9810843-0-6,
ISBN-13: 978-3-9810843-0-6, pp. 119-224, 2006.

© 2006 by John von Neumann Institute for Computing
Permission to make digital or hard copies of portions of this work for personal or classroom use is granted provided that the copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise requires prior specific permission by the publisher mentioned above.

http://www.fz-juelich.de/nic-series/volume34

Modeling of Water Explicitly in the Replica-Exchange Simulation Method for Protein Folding

Mu Yuguang* **and Yang Ye**

School of Biological Sciences,
Nanyang Technological University, Singapore 637551

*Correspondent Author: *E-mail: ygmu@ntu.edu.sg*

Considering the replica-exchange simulation of protein in explicit water to be a two-Hamiltonian system with one Hamiltonian for conformational sampling in molecular dynamics simulation and another for controlling the exchanges between difference temperature replicas in Monte Carlo jumps, we introduce an approximation on the latter Hamiltonian using a continuum solvent model, surface-generalized Born model. Such replica exchange simulation with hybrid Hamiltonians method is applied to fold the C-terminus (residue 41-56) of protein G from extended structures. Promising results show that not only the total number of replica needed is largely reduced but also the folding efficiency is greatly enhanced. Combined with recently invented dihedral principle component analysis a general framework for *ab initio* folding a small protein merely from sequence knowledge is emerging.

PACS numbers: 87.15.Cc, 87.15.Aa

Predicting protein structure solely from its amino acid sequence has long been a great challenge in modern molecular biology[1,2]. Although remarkable progress was made recently by Baker and his collaborators[3,4], the problem remains unsolved. Recent studies on peptides[5,6] by molecular dynamics simulation tools using explicit water model, aided by advanced sampling strategies, such as replica exchange molecular dynamics (REMD) method[7], show a prospective way towards solving the folding puzzle. However when REMD is applied to larger protein system a huge computation facility will be needed because the number of replicas needed increases simultaneously with the increase of the degrees of freedom of the system[8]. Endeavors to overcome the shortage of REMD have been undertook by several groups[8–12].

The Hamiltonian of a protein solvated in an aqueous environment can be written as $H_{sys} = H_p + H_{pw} + H_w$, where p, pw and w denote protein-protein, protein-water and water-water interactions respectively. In standard REMD scheme, replicas are propagating independently at certain temperatures with Monte Carlo (MC) exchange at certain intervals. The MC exchange corresponds to the Metropolis criterion following the detailed balance condition: $p_{i\leftrightarrow j} = min\left(1, e^{\Delta\beta\Delta E}\right)$, where $\Delta\beta = \beta_j - \beta_i$, $\beta = \frac{1}{k_B T}$, $\Delta E = E\left(\chi_j - \chi_i\right)$ and χ_i, χ_i are the configurations of the neighboring replicas. The total energy, $E\left(\chi\right)$, should be the same as the system Hamiltonian and can also be decomposed into three terms: $E\left(\chi\right) = E_p + E_{pw} + E_w$ as well. In explicit water REMD simulation many water molecules are employed to solvate well the unfolded protein. This makes the total energy, E, and its differenceds between neighbours, ΔE, huge. In order to get a reasonable high exchange probability between the neighboring temperature ladders, $p_{i\leftrightarrow j}$, the temperature jumps, $\Delta\beta$, should be small. Therefore a large number of replicas are needed to cover a broad temperature range, which from $300K$ to $600K$. Realizing that

the water-water interaction makes the largest contribution to the total energy a possible workaround could be that decouple the water-water interaction from the thermal baths and make the ΔE small though the total energy E nearly unchanged. Recent works done by Berne group[9] and Simmerling group[11] were pursued in this direction using different strategies. However such decoupling not only causes unphysical Hamiltonian on most replicas in MD simulation, the resulting hot protein and cold water effect would also hinder the conformation sampling efficiency at high temperatures.

This letter presents an improved solution to this problem. The standard REMD method can be thought as a hybrid method of MD and MC simulation. The Hamiltonians for both MD and MC should be the same for the sake of consistency. However if one checks the roles of both Hamiltonians in details, such consistency is found to be not a necessity. The full system Hamiltonian, H_{sys}, is necessary for MD simulations to correctly sample conformational space. While for MC steps we find it is not necessary to use the same Hamiltonian, H_{sys}. In H_{sys} the water-water interaction is dominant but is not directly related to the protein folding. The key of the MC steps in REMD is to help protein jump out of its local minima. Evidently the REMD could be more efficient in folding protein if the Hamiltonian for MC steps represents the energies of the protein more directly. In light of that we introduce an approximation to the Hamiltonian for MC steps. Instead of using the H_{sys}as in the MD steps a second Hamiltonian with the same energies for protein-protein interaction but protein-water and water-water interactions approximated by a continuum solvent model, surface-generalized Born model (GBSA)[13] is used in MC steps. In this way the dominate water-water interaction is averaged out and therefore a small ΔE is obtained. At current moment the implicit water model or continuum solvent model, although very computationally efficient, was found not to be able to fold peptide correctly[14]. Our method is a complement for such deficiency in the GBSA model.

We call this new version of REMD as REMD with hybrid Hamiltonians (REMDhH). Hybrid Hamiltonians costs the violation of detailed balance of the whole system to some degree. Considering the interested temperature range for REMD is from $300K$ to $600K$ the water behaves well as a liquid (in constant volume REMD) and should be quickly relaxed to its equilibrium state when the system jumps to a new temperature.

Here REMDhH is applied to fold the C terminus (residue 41-56) of protein G, a 16 amino acids peptide which was found to be able to fold into a beta-hairpin *in vitro*[15]. The GROMACS program suite[16] and the full atomic OPLS-AA force field[17] are used. The peptide is capped with the normal ACE and NME groups with 256 atoms in total. It is solvated by 5469 water molecules plus three K+ ions to neutralize the charged molecular system. The whole system consists of 16666 atoms in a cubic simulation box of 5.3 nm length. Sixteen replicas are used whose temperatures are 300.0, 317.2, 335.4, 354.6, 374.9, 396.4, 419.1, 443.1, 468.5, 495.4, 523.8, 553.8, 585.5, 619.1, 654.5 and 692.1 Kelvin. While 64 replicas had been used for the same peptide with less water molecules (1361) studied by standard REMD[5]. The GBSA energies are calculated using Tinker program[18]. The interface between GROMACS and Tinker is built by modifying GROMACS source code. A twin-range cutoff of 0.9/1.4 nm is used for the non-bonded interactions and a reaction-field correction with permittivity is employed. The integration

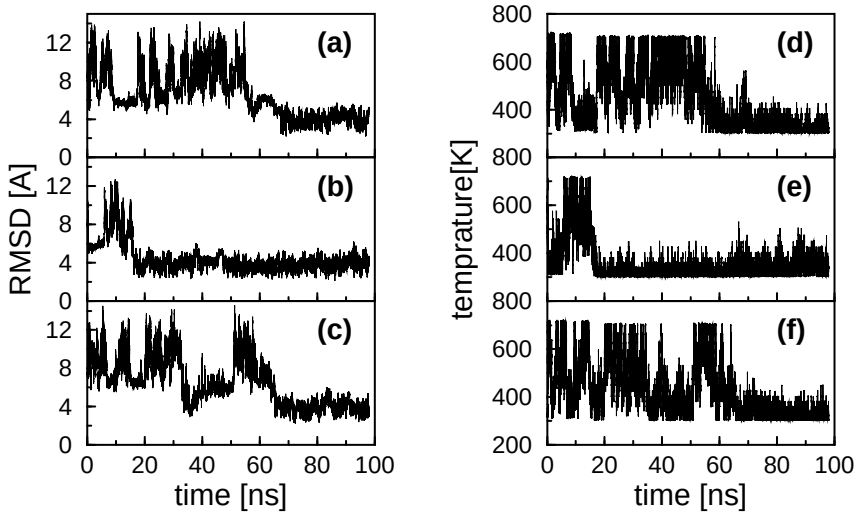

Figure 1. (a), (b) and (c) are the three folded replica trajectories of RMSD from NMR structure, (d), (e) and (f) are the trajectories of temperature jumps of the related replicas, respectively.

step in all simulations is 0.002 ps. Non-bonded pair lists are updated every 10 integration steps. The system is coupled to an external heat bath with a relaxation time of 0.1 ps. All bonds including hydrogen atoms are constrained in length. Each replica is run for 98 ns with replica exchange attempted every 2 ps. The final acceptance ration for replica exchanges is found to be 35%-43%.

To test the validity of the method the initial configurations of peptide are extended structures, which means only the information of amino acids sequence is supplied. Three folding events are found during the simulation. Figure 1a, 1b and 1c show the trajectories of root mean squared deviations (RMSD) from the NMR structure[19] calculated with all atoms. The fastest folding happens in replica 9 (Figure 1b) around simulation time of 18 ns. The following folding events happen in replica 4 and 15 (Figure 1a and 1c) around simulation time of 60 ns. All the folded structures maintain stable until the end of simulations. Here we denote a structure as folded when its RMSD drops below 4A. A more detailed analysis is given below. The folding processes revealed by REMD are a process of annealing and relaxation. The temperature trajectories (Figure 1d, 1e and 1f) of the three folded replicas indicate this view clearly. Accompanying with the decrease of RMSDs the temperatures are cooling down.

In order to find the quantitative reasons why the employed REMDhH method is superior to the standard REMD, the correlation between RMSDs and the total energy of the system with explicit water molecules, and the correlation between RMSDs and the GBSA energies are calculated and displayed in Figure 2a and 2b respectively. It is clearly shown that the correlation between the RMSDs and the total energies used by regular REMD is poor. Its correlation coefficient is 0.01. While the correlation between

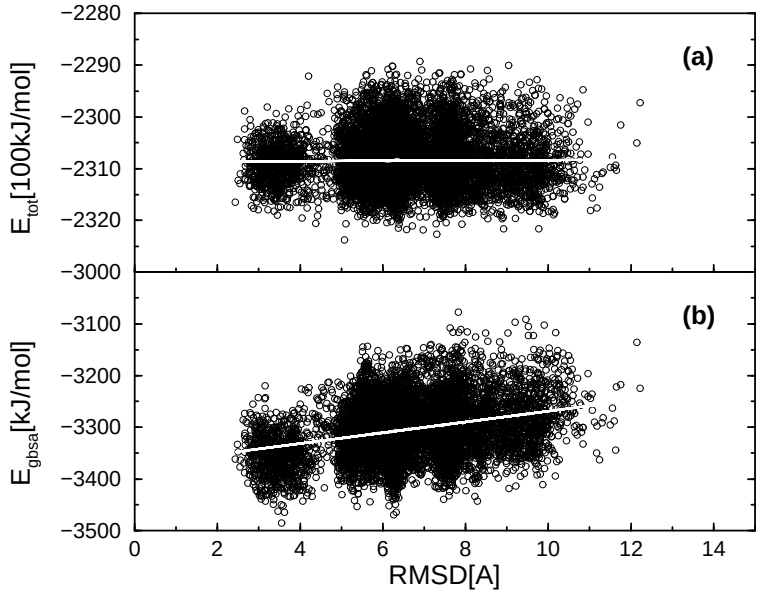

Figure 2. (a) Correlations of the total energies, including protein-protein, protein-water, and water-water interactions, with RMSD from NMR structure. The correlation coefficient is 0.01. (b) Correlations of the GBSA energies with RMSD. The correlation coefficient is 0.3. The white lines are the linear regression results.

the RMSDs and the energies from the GBSA model is much stronger whose correlation coefficient is 0.3. As mentioned earlier the folding process in REMD method could be regarded as an annealing process. To perform such annealing efficiently two factors for the MC steps must be considered. First the lower energy conformations could be identified in the high temperature conformational ensemble. Secondly these lower energy conformations have to be quickly annealed to the lower temperature simulations. The first factor requires that the energies for MC steps should interrogate the protein conformation directly. Otherwise the lower energy conformations would not easily be identified. The second factor claims that the fewer temperature ladders the quicker the annealing process which implies that the number of replicas in REMD should be minimized. The implementation of the current REMDhH could make both requirements satisfactory. If the REMD efficiency is approximated to be linearly correlated with the correlation coefficient of RSMD/energy for MC steps the speedup of folding by REMDhH compared with that of standard REMD is a factor of 30 for this system.

In protein structure prediction usually the folded structure is unknown. The RMSDs from the native structure and the fraction of native contacts which are commonly used as folding reaction coordinates are not available in such situations. In order to identify the global minimum new suitable reaction coordinates are desirable. We find that the recently invented dihedral principal component analysis (dPCA) method works[20] fine here. Figure 3 reveals the folding free energy curve for this beta-hairpin peptide obtained by projecting conformational ensemble of $T = 300K$ onto the first eigenvector of dPCA

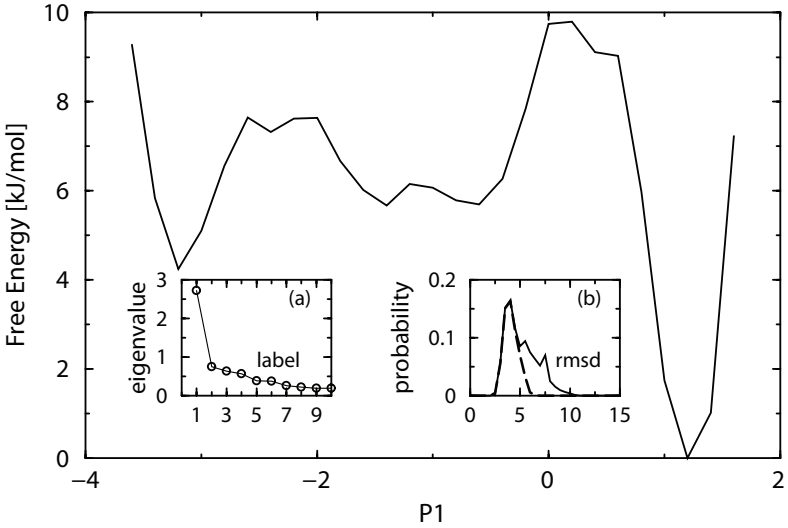

Figure 3. Folding free energy curve of the GB1 hairpin peptide obtained by projection onto the eigenvector of largest eigenvalue using dihedral PCA method. Inset (a) shows the first 10 largest eigenvalues of dPCA result. Inset (b) shows the distribution of RMSD of ensemble at $T = 300K$ (solid line) and the distribution of RMSD of the ensemble within the global minimum using such criteria $P1 > 0.7$ (dashed line).

whose eigenvalue is shown to be much larger than other eigenvalues in inset (a). The free energy curve is funnel-like and it is easy to identify the global minimum which is located around $P1 = 1.2$. Here $P1$ is the projection value on the first eigenvector. To check whether this global minimum is related to the native state the distribution of RMSDs from this ensemble is plotted in Figure 3 inset (b) by the dashed line. For comparison the distribution of RMSDs for all the structures obtained at $300K$ is shown by the solid line. Evidently the structures from the global minimum are native-like with average $RMSD = 3.5A$. And $P1$ is a better reaction coordinate than RMSD in the sense that the former can distinguish folded states from unfolded ones more clearly.

In summary a new version of REMD method is suggested. By decomposing REMD into two steps, one is MD step and the other is MC step, different Hamiltonians can be applied to them. In MD step the full Hamiltonian with all protein-protein, water-water, water-protein interactions included guarantees that the conformations of protein are sampled correctly. In MC step, however, only the protein-protein interaction is explicitly considered, the water-water and water-protein interactions are approximated by a continuum GBSA model. Such a hybrid Hamiltonian scheme reduces the number of replica greatly and meanwhile increases the folding efficiency by more than a factor of 10. On the other hand, dPCA analysis makes the global minimum easily identified if protein is folded in simulation. Combining REMDhH with dPCA a genenal framework for ab initio protein structure prediction is emerging. Together with the available of accurate force fields and powerful computational facility the era of solving protein folding by brute force MD simulations will be coming soon[21].

Acknowledgments

The support of a Lee Kuan Yew Research Fellowship to Y.G.M. is acknowledged. The support of PhD study in Nanyang Technological Univerisity to Yang Ye is appreciated. The simulations were performed on the supercomputer of Bioinformation Research Center and the Frankfurt Center of Scientific Computing, which are acknowledged by the generous allocation of CPU time.

References

1. D. J. Wales and H. A. Scheraga, Science **285**, 1368 (1999).
2. J. N. Onuchic and P. G. Wolynes, Curr. Opin. Struct. Biol. **14**, 70 (2004).
3. K. T. Simons, C. Kooperberg, E. Huang, and D. Baker, J. Mol. Biol. **268**, 209 (1997).
4. C. A.Rohl, C. E.Strauss, K. M.Misura, and D. Baker, Methods Enzymol **383**, 66 (2004).
5. R. Zhou, B. J. Berne, and R. Germain, Proc. Natl. Acad. Sci. USA **98**, 14931 (2001).
6. A. E. Garcia and J. N. Onuchic, Proc. Natl. Acad. Sci. USA **100**, 13898 (2003).
7. Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **314**, 141 (1999).
8. H. Fukunishi, O. Watanabe, and S. Takada, J. Chem. Phys. **116**, 9058 (2002).
9. P. Liu, B. Kim, R. A. Friesner, and B. J. Berne, Proc. Natl. Acad. Sci. USA **102**, 13749 (2005).
10. T. Z. Lwin and R. Luo, J. Chem. Phys. **123**, 194904 (2005).
11. X. Cheng, G. Cui, V. Hornak, and C. Simmerling, J. Phys. Chem. B **109**, 8220 (2005).
12. E. Lyman, F. M. Ytreberg, and D. M. Zuckerman, Phys. Rev. Lett. **96**, 028105 (2006).
13. Q. Liu, P. S. Shenkin, F. P. Hollinger, and W. C. Still, J. Phys. Chem. A **101**, 3005 (1997).
14. R. Zhou and B. J. Berne, Proc. Natl. Acad. Sci. USA **99**, 12777 (2002).
15. F. J. Blanco, G. Rivas, and L. Serrano, Nat. Struct. Biol. **1**, 584 (1994).
16. H. J. C. Berendsen, D. van der Spoel, and R. van Drunen, Comp. Phys. Comm. **91**, 43 (1995).
17. W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, J. Am. Chem. Soc. **118**, 11225 (1996).
18. R. V. Pappu, R. K. Hart, and J. W. Ponder, J. Phys. Chem. B **102**, 9725 (1998).
19. A. M. Gronenborn, D. R. Filpula, N. Z. Essig, A. Achari, M. Whitlow, P. T. Wingfield, and G. M. Clore, Science **253**, 657 (1991).
20. Y. G. Mu, P. H. Nguyen, and G. Stock, Proteins **58**, 45 (2005).
21. M. Karplus and J. A. McCammon, Nat. Struct. Biol. **9**, 646 (2002).