Protein Folding Simulations from the First Principles

Group Representative
Yuko Okamoto Institute for Molecular Science

Authors
Yuko Okamoto Institute for Molecular Science
Yuji Sugita Institute of Molecular and Cellular Biosciences, University of Tokyo
Takao Yoda Nagahama Institute of Bio-Science and Technology
Ayori Mitsutake Faculty of Science and Engineering, Keio University
Takeshi Nishikawa National Institute of Advanced Industrial Science and Technology

It is one of the most challenging problems in computational bioscience to predict three-dimensional structures of proteins with the input of only the amino-acid sequence information (prediction from the first principles). The goal of the present project is to succeed in the prediction of the three-dimensional structures of a small protein from the first principles. For this purpose, we continued our efforts to tune the codes of our replica-exchange molecular dynamics program REMD so that it will perform optimally on the Earth Simulator. While our tuning was quite limited last year (the vectorization ratio of 96.51%, the parallelization ratio of 92.88%, and the parallelization efficiency ratio of 80.98% for the system of a small protein in explicit water with only eight nodes of the Earth Simulator) because we were allowed only a short period (3 months), this year we could achieve the vectorization ratio of 97.77%, the parallelization ratio of 99.96%, and the parallelization efficiency ratio of 73.70% for the same system with as many as 112 nodes of the Earth Simulator. We started a production run of folding simulation of a small protein with 56 amino acids.

Keywords: Protein Structure Predictions, Protein Folding Problem, Molecular Dynamics, Generalized-Ensemble Algorithms, Replica-Exchange Method

Report of the Results

There is a close relationship between the three-dimensional structures of proteins and their biological functions. The study of protein structures is thus aimed at the understanding of how proteins carry out their functions. The research in this field is ultimately led not only to drug design and *de novo* design of artificial proteins with specific functions but also the elucidation of the pathogenic mechanism for the disease that is caused by misfolding of proteins (such as mad cow disease and Alzheimer's disease).

It is widely believed that the three-dimensional structures of proteins are determined solely by their amino-acid sequence information. However, the prediction of protein structures by computer simulations with the input of only the amino-acid sequence (prediction from the first principles) has yet to be accomplished. The main difficulty lies in the fact that the number of internal degrees of freedom of protein systems is extremely large, and there exist a huge number of local minima in the energy function. It is a very challenging problem to find the global-minimum state in free energy, which corresponds to the native protein structure, because simulations by conventional algorithms will get trapped in one of the local-minimum states. In order to overcome this difficulty, we have developed three powerful simulation methods (which are examples of generalized-ensemble algorithms; for a review, see Ref.[1)]). They are replica-exchange molecular dynamics (REMD)[2)], replica-exchange multicanonical algorithm (REMUCA)[3)–5)], and multicanonical replica-exchange method (MUCAREM)[3)–5)]. The first method, REMD, has been immediately accepted by the protein folding community as soon as we announced it in Ref. 1), and REMD is now employed by the IBM BlueGene Project[6)] and is also incorporated into a standard program package, AMBER version 8,[7)] for protein simulations.

The goal of the present project is to succeed in the prediction of the three-dimensional structures of proteins from the first principles by employing the powerful simulation algorithms that we developed (namely, REMD, REMUCA, and MUCAREM). In particular, we try to predict, for the first time, the three-dimensional structure of a small protein with about 50 amino acids in water by simulations with atomistic details incorporated.

This year we have concentrated on the tuning of the source code REMD, which incorporates the above three algorithms, REMD, REMUCA, and MUCAREM, so that it can achieve optimal performances on the Earth Simulator. The tuning was carried out by taking the system of a small protein, protein G, in explicit water and using up to 112 nodes of the Earth Simulator. The source code is a molecular dynamics code that is based on generalized ensemble. The system consists of a protein of 56 amino acids that is placed in a sphere of water molecules with radius 50 angstroms. The total number of water molecules is 17,187 and the total number of atoms in the entire system of protein and water is 52,416 (see Fig. 1).

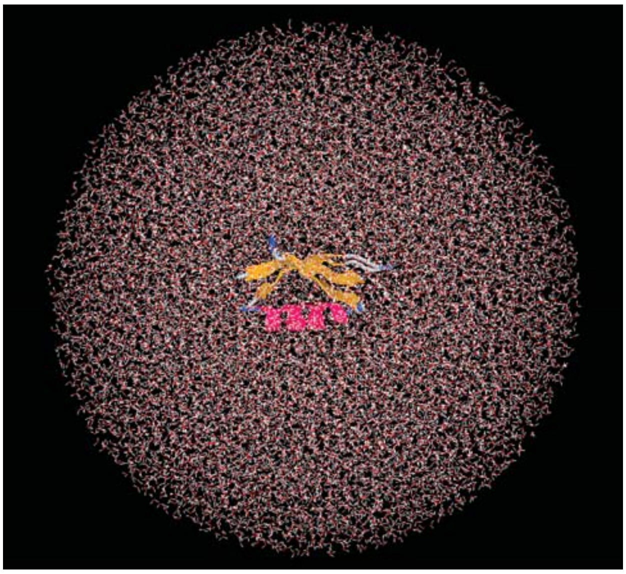

Fig. 1 Protein G in a sphere of water of radius 50 angstroms. The numer of amino acids is 56, the number of atoms in the protein is 855, and the total number of atoms of the entire system is 52,416.

While our tuning was quite limited last year (the vectorization ratio of 96.51%, the parallelization ratio of 92.88%, and the parallelization efficiency ratio of 80.98% for the system of a small protein in explicit water with only eight nodes of the Earth Simulator) because we were allowed only a short period (3 months), this year we could achieve the vectorization ratio of 97.77%, the parallelization ratio of 99.96%, and the parallelization efficiency ratio of 73.70% for this system, using as many as 112 nodes of the Earth Simulator. We ran replica-exchange MD simulations with 224 replicas and confirmed that replica-exchange simulations are properly performing even with this many replicas.

We think that we are finally ready for making production runs of our protein folding simulations on the Earth Simulator.

References

1) A. Mitsutake, Y. Sugita, and Y. Okamoto, "Generalized-ensemble algorithms for molecular simulations of biopolymers," Biopolymers (Peptide Science), vol. 60, no. 2, pp. 96–123, August 2001.
2) Y. Sugita and Y. Okamoto, "Replica-exchange molecular dynamics method for protein folding," Chemical Physics Letters, vol. 314, nos. 1–2, pp. 141–151, November 1999.
3) Y. Sugita and Y. Okamoto, "Replica-exchange multicanonical algorithm and multicanonical replica-exchange method for simulating systems with rough energy landscape," Chemical Physics Letters, vol. 329, nos. 3–4, pp. 261–270, October 2000.
4) A. Mitsutake, Y. Sugita, and Y. Okamoto, "Replica-exchange multicanonical algorithm and multicanonical replica-exchange Monte Carlo simulations of peptides. I. Formulation and benchmark test," Journal of Chemical Physics, vol. 118, no. 14, pp. 6664–6675, April 2003.
5) A. Mitsutake, Y. Sugita, and Y. Okamoto, "Replica-exchange multicanonical algorithm and multicanonical replica-exchange Monte Carlo simulations of peptides. II. Application to a more complex system," Journal of Chemical Physics, vol. 118, no. 14, pp. 6676-6688, April 2003.
6) http://www.research.ibm.com/bluegene/
7) http://amber.scripps.edu/