Protein Folding Simulations from the First Principles

Project Representative
Yuko Okamoto	Department of Physics, Nagoya University

Authors
Yuko Okamoto	Department of Physics, Nagoya University
Yuji Sugita	Theoretical Biochemistry Laboratory, RIKEN
Takao Yoda	Nagahama Institute of Bio-Science and Technology
Ayori Mitsutake	Faculty of Science and Engineering, Keio University
Takeshi Nishikawa	Global Scientific Information and Computing Center, Tokyo Institute of Technology
Yoshitake Sakae	Department of Physics, Nagoya University

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 chose a small protein with 56 amino acids (B1 domain of streptococcal protein G). We first performed a replica-exchange molecular dynamics (REMD) simulation of protein G in vacuum with 96 replicas. The initial conformation was a fully extended one. We solvated one of the obtained compact conformations in a sphere of water of radius 50 angstroms. The total number of water molecules was 17,187 (the total number of atoms was 52,416 including the protein atoms). Using 112 nodes of the Earth Simulator, we performed a REMD simulation of this system with 224 replicas. The REMD simulation was successful in the sense that we observed a random walk in potential energy space that suggests that a wide conformational space was sampled. We used the results of this REMD simulation to prepare for an even more powerful generalized-ensemble algorithm, namely, multicanonical replica-exchange method (MUCAREM). Using 112 nodes of the Earth Simulator, we succeeded in performing a MUCAREM simulation of this system with 56 replicas. By analyzing the results of these simulations, we found that a structure very similar to the native one has been obtained.

Keywords: Protein structure predictions, protein folding problem, molecular dynamics, generalized-ensemble algorithms, replica-exchange method, multicanonical 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 the local-minimum states. In order to overcome this difficulty, we have developed two powerful simulation methods (which are examples of generalized-ensemble algorithms; for a review, see Ref.[1]). They are replica-exchange molecular dynamics (REMD)[2] 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 and is also incorporated into standard program packages such as AMBER, CHARMM, NAMD, GROMACS, etc. 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 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 continued and extended the molecular dynamics simulations based on one of the generalized-ensemble algorithms, namely, REMD, using up to 112 nodes of the Earth Simulator. The system that we studied is a small protein, protein G, with 56 amino acids. The total number of atoms in the protein is 855. We first performed a REMD simulation of protein G in vacuum with 96 replicas. The initial conformation of the REMD simulation was a fully extended one. We then solvated one of the obtained compact conformations in a sphere of water of radius 50 angstromes. The total number of water molecules was 17,187 (the total number of atoms was 52,416 including the protein atoms). In Fig. 1 we show the native structure of this protein, which was obtained by X-ray crystallographic experiments, solvated in the sphere of water (our goal).

Using 112 nodes of the Earth Simulator, we performed a REMD simulation of this system with 224 replicas. The REMD simulation was successful in the sense that we observed a random walk in potential energy space that suggests that a wide conformational space was sampled.

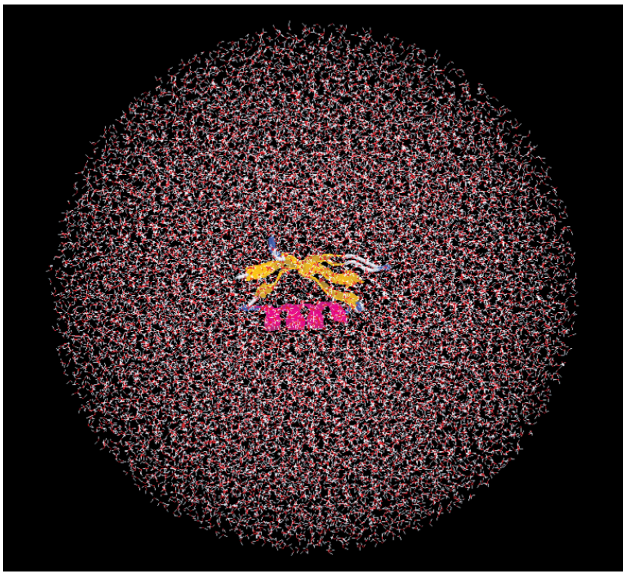

Fig. 1 Native structure of protein G solvated in a sphere of water of radius 50 angstromes.

In Fig. 2 we show the canonical probability distributions of the total potential energy at the corresponding 224 temperatures ranging from 250 K to 700 K. As is clear from the Figure, all the adjacent distributions have sufficient overlaps with the neighboring ones, indicating that this REMD simulation was successful.

We indeed observed a random walk in the potential energy space. This random walk in potential energy space induced a random walk in the conformational space, and we indeed observed many occasions of the formation of native-like secondary structures (α-helices and β-strands) during the REMD simulation.

In Fig. 3 we show the time series of the potential energy. We indeed observe a random walk in potential energy space, which is expected for a properly performed REMD simulation.

Using the results of this REMD simulation, we then performed a MUCAREM simulation, which is more powerful than REMD. In Fig. 4 we show the probability distributions of the total potential energy at the corresponding 56 multicanonical ensembles. As is clear from the Figure, all the adjacent distributions have sufficient overlaps with the neighboring ones, indicating that this MUCAREM simulation was successful.

Analyzing the results of these REMD and MUCAREM simulations, we found that a structure very similar to the native one has been obtained. In Fig. 5 we compare this structure with the native one. The overall structure is indeed very similar.

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

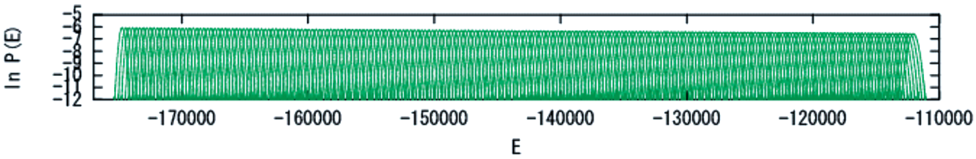

Fig. 2 The canonical probability distributions of the total potential energy of protein G obtained from the REMD simulation with 224 temperatures. They are all bell-shaped with sufficient overlaps with the neighboring ones.

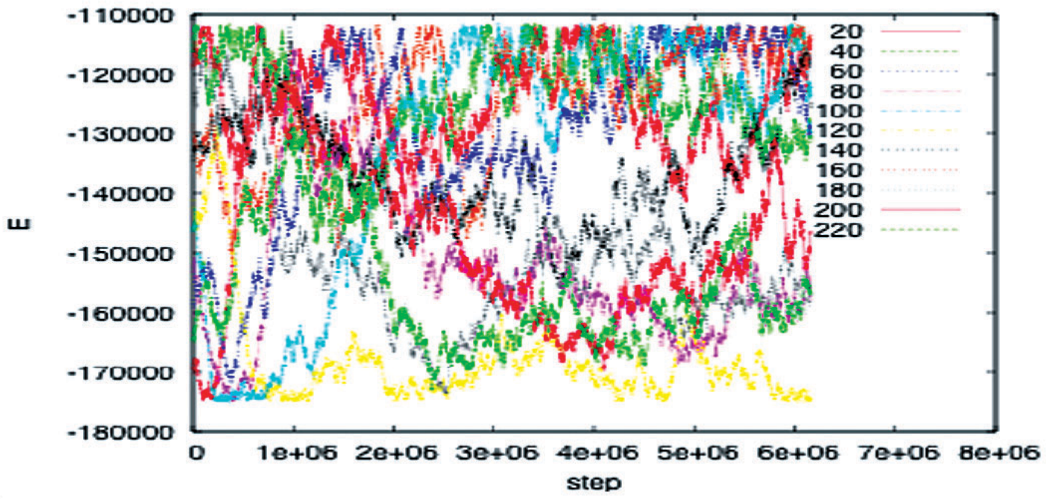

Fig. 3 Time series of potential energy of some of the replicas during the REMD simulation of protein G in explicit solvent.

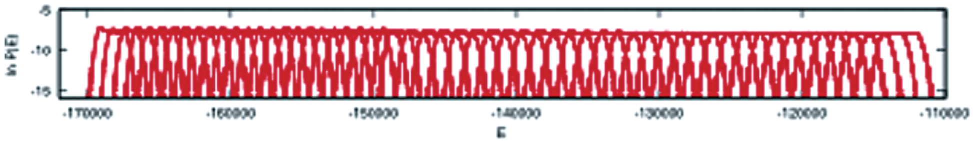

Fig. 4 The probability distributions of the total potential energy of protein G obtained from the MUCAREM simulation with 56 replicas. They all have sufficient overlaps with the neighboring ones.

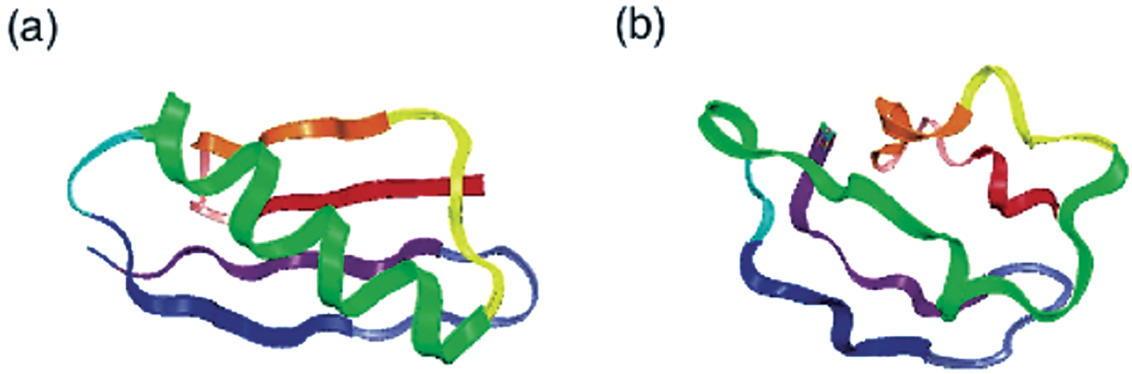

Fig. 5 (a) The native structure of protein G and (b) a structure that was obtained by the REMD and MUCAREM simulations of protein G in explicit solvent.

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.

第一原理からのタンパク質の折り畳みシミュレーション

プロジェクト責任者
岡本　祐幸　　名古屋大学　大学院理学研究科

著者
岡本　祐幸　　名古屋大学　大学院理学研究科
杉田　有治　　理化学研究所　理論生化学研究室
依田　隆夫　　長浜バイオ大学　バイオサイエンス学部
光武亜代理　　慶應義塾大学　理工学部
西川　武志　　東京工業大学　学術国際情報研究センター
榮　　慶丈　　名古屋大学　大学院理学研究科

1960年代初頭のアンフィンゼンの実験以来、タンパク質の自然の立体構造は、アミノ酸配列の情報及び周りの溶媒環境のみで決まっており、自由エネルギーの最小状態に対応すると広く信じられている。しかし、系にエネルギー極小状態が無数に存在するために、一定温度のモンテカルロ法や分子動力学法等による従来のシミュレーションでは、それら極小状態の近傍に留まってしまって、立体構造予測シミュレーションが絶望的に難しくなる。本研究の目的はこの困難を拡張アンサンブル法を適用することによって克服し、水分子をあらわに取り入れた分子シミュレーションによって、小タンパク質の折り畳みに成功することである。我々はアミノ酸数56個の小タンパク質であるProtein Gにおいて、レプリカ交換分子動力学法(REMD)によるシミュレーションを地球シミュレータ上で実行している。このタンパク質は原子数が855個である。まず、真空中で初期構造として完全に伸びた構造から96レプリカのREMDシミュレーションを実行し、得られたコンパクトな構造をもつProtein Gを半径50Åの水球中(水分子の数は17,187個)に入れて、全体として、原子数が52,416個の系を考慮した。一昨年、我々は、この系において、地球シミュレータ112ノードを用い、レプリカ数が224のREMDシミュレーションを実行に成功した。今年度は更に、データを蓄積した。エネルギー空間上のランダムウォークが得られ、REMDシミュレーションが成功したと言える。タンパク質系においてはこれほど大規模の系におけるレプリカ交換シミュレーションの成功は初めてのことである。そして、この結果を使って、より強力な拡張アンサンブル法であるマルチカノニカルレプリカ交換法(MUCAREM)のシミュレーションの準備をした。そして、地球シミュレータ112ノードを用い、レプリカ数56のMUCAREMシミュレーションの実行にも成功した。そして、その結果を解析したところ、自然の構造に似た構造が得られた。

キーワード: タンパク質の立体構造予測, タンパク質の折り畳み問題, 分子動力学シミュレーション, 拡張アンサンブル法, レプリカ交換法, マルチカノニカルレプリカ交換