Studies on Protein Dynamical Structures and Electronic States by Large-Scale Computations

Group Representative
Fumitoshi Sato Institute of Industrial Science, University of Tokyo, Visiting Associate Professor

Authors
A. Yuko Okamoto Institute for Molecular Science, Associate Professor
B. Minoru Saito Faculty of Science and Technology, Hirosaki University, Professor
C. Fumitoshi Sato Institute of Industrial Science, University of Tokyo, Visiting Associate Professor

Protein is a huge molecule connected with many amino acid residues. It forms characteristic structure, changes its structure every moment, and advances a peculiar reaction with delivering and receiving electrons. In order to understand the function of protein, the research was promoted by the following three subgroups.

A; "Protein Folding Simulations from the First Principles",

B; "Realistic Simulations for the Structural Changes of Proteins",

C; "All-electron Calculation on Very Large-Sized Proteins by Density Functional Method".

In this year, all subgroups made effort to install and tune the codes of programs developed by them to the Earth Simulator. In the subgroup A, the program REMD, which can search for the global stable structure of proteins, achieved the vectorization ratio of 96.51 %, the parallelization ratio of 92.88 %, and the parallelization efficiency ratio of 80.98 % under the system of a small protein in explicit water. The purpose of subgroup B is to computationally visualize the structural changes of proteins using COSMOS90 which can efficiently simulate proteins in water with all degrees of freedom and long-range Coulomb interactions. It was successfully vectorized (ratio is 98% and then acceleration is 8.5 times). ProteinDF is the C++ program of the subgroup C, and can calculate the all-electron wavefunction of proteins. The core routine was vectorized with 92% of efficiency. In addition, matrix operation routines were almost transposed to those in ASL/ES library.

In the next year, we will achieve further acceleration and begin own calculations with the concrete system.

Keywords: Protein, Protein Folding Problem, Generalized-Ensemble Algorithms, Protein Structural Change, Molecular Dynamics, All-Electron Calculation on Protein, Density Functional Method

Report of the result:

Subgroup A; "Protein Folding Simulations from the First Principles"

Subgroup A this year has concentrated on the tuning of the source code REMD, which will be used in the present project, 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 one node 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 35 angstroms. The total number of atoms in the system is 17,784. We have reached the vectorization ratio of 96.51 %, the parallelization ratio of 92.88 %, and the parallelization efficiency ratio of 80.98 %. We ran replica-exchange MD simulations with 4 replicas. The following are some of the details of outputs from our tuning.

```
*----------------------------------------*
  FLOW TRACE ANALYSIS LIST
*----------------------------------------*
  Execution : Fri Feb 28 17:27:22 2003
  Total CPU : 0:08'26"023
  PROG.UNIT        EXCLUSIVE      AVER.TIME
MOPS               MFLOPS         V.OP
                   TIME[sec](%)   [msec]
                   RATIO
  dnc151x          233.891(46.2)   467.781
5817.4             2001.0         99.64
  cutcmmx          126.372(25.0)  2106.204
345.1                0.0          18.73
  cmmepcx           49.751(9.8)     99.503
9902.9             4498.5         99.73
       ..............
  Total            503.820(100.0)    0.023
3888.4             1413.0         96.51
```

We expect further improvement (both in vectorization ratio and parallelization ratio) with the actual system that we are going to simulate in which we will have about 100,000 atoms instead of the present case of 17,784 atoms.

Subgroup B; "Realistic Simulations for the Structural Changes of Proteins"

Proteins thermally fluctuate in the water environment. They largely change their structures to undergo the biologically functions. For example, a hemoglobin molecule can efficiently transfer oxygen molecules from the lungs to the muscles. The binding of an oxygen molecule enhances additional oxygen bindings to other sites. Various experimental studies revealed that this cooperative binding is associated with large structural change. However, the experimental studies could not reveal the dynamical features of the structural changes, although they observed the structural difference between the initial and final states. The purpose of our group is to visualize computationally such structural changes of proteins using the Earth Simulator and a software COSMOS90. COSMOS90 was developed by one of the authors (M.S.) to efficiently simulate the protein in water with all degrees of freedom and long-range Coulomb interactions. To achieve our purpose we need both, because of the long-time simulation for the large size of the molecular system (10^5 atoms and 10^9 steps). COSMOS90 was developed on the vector-parallel machines (VPP500 and VPP5000) about ten years ago. Recently, we successfully installed it on the Earth Simulator using MPI as a joined project between our group and the Earth Simulator center (from January to March in 2003).

Table 1 Execution time (sec) for 1step MD simulation of a protein in water. *Estimation

Machine/ No. of PE	Scalar	Vector
SR2201/		
1	21.88	-
128	0.30	-
256	0.20	-
VPP500/		
1	9.254	0.886
4	2.382	0.233
8	1.279	0.125
VPP5000/		
1	1.801	0.146
4	0.457	0.040
8	0.233	0.023
10	0.189	0.018
E.Simulator		
/ 1	1.593	0.189
1 node	-	0.027
	-	0.029*
8 node	-	0.008
	-	0.005*

The performance speed of COSMOS90 on the Earth Simulator was shown in Table I together with the speed on VPP500 and VPP5000, where the speed is defined as the time to execute a step of simulation for DNA and protein in water (16034 atoms, Fig.1). Speed with the sign * means the expectant speed estimated from the difference in the hardware between VPP5000 and the Earth Simulator. The vector speed of a single processor 0.189 is 8.5 times faster than the scalar speed 1.593. This result means that COSMOS90 is well vectorized on the Earth Simulator (vector ratio is 98%). The vector speed of a single node 0.027 is almost same as the expectant speed (0.029). The speed of 8 nodes (consisting of 64 processors) 0.008 is faster than that of VPP5000 but still slower than the expectant speed (0.005) of the Earth Simulator. Further speed up by parallelization is the purpose of our group in the next stage (from April to September in 2003).

Subgroup C: "All-electron Calculation on Very Large-Sized Proteins by Density Functional Method"

This subgroup developed the quantum chemical calculation software ProteinDF for large-scale proteins by the density functional method, and succeeded in the calculation of all-electron wavefunction of 104 residues metal protein cytochrome *c* (9,600 orbitals) with a 15 workstation cluster (theoretical peak = 15GFLOPS) for the first time. Now, this is only software that can perform all-electron calculation on 100 residues metalloproteins, including electron correlation. From this research, it is estimated that all-electron calculation on 100,000 orbitals protein is possible with double precision. This is equivalent to the 1,000 residues proteins, and almost all important proteins become the candidates for calculation. The purpose of this subgroup applies this ProteinDF to the Earth Simulator and carried out all-electron calculation on 1,000 residues important protein in the three years (2003-2005).

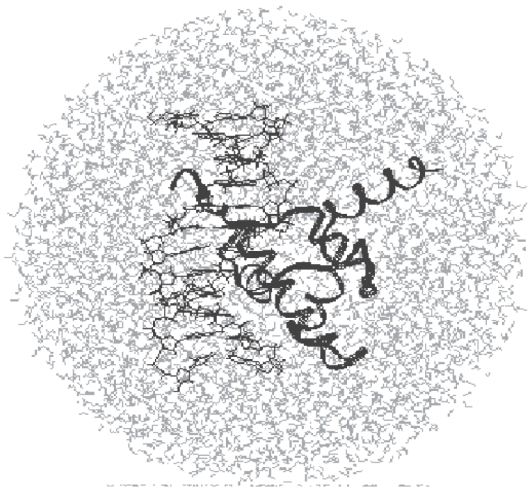

Fig. 1 Complex of DNA and protein in water (16,034 atoms).

ProteinDF is coded by C++. It was successfully installed on the Earth Simulator. ProteinDF consists of four time-consuming routines; molecular integrals, exchange correlation fitting, diagonalization and the other matrix operations. Their tasks are depending on the 2.3rd, 1.8th, 3.3rd and 2.9th power of the number of orbitals, respectively (Fig. 2). In this year, molecular integral routine was vectorized with 92% of efficiency. In addition, all matrix operations including diagonalization were almost transposed to those in ASL/ES library. ProteinDF was already parallelized, and the efficiency is 85% with 100BaseTX network. Then, in the next year, we will achieve further acceleration including exchange correlation fitting routine, and calculate the all-electron wavefunction of 200 - 300 residues protein.

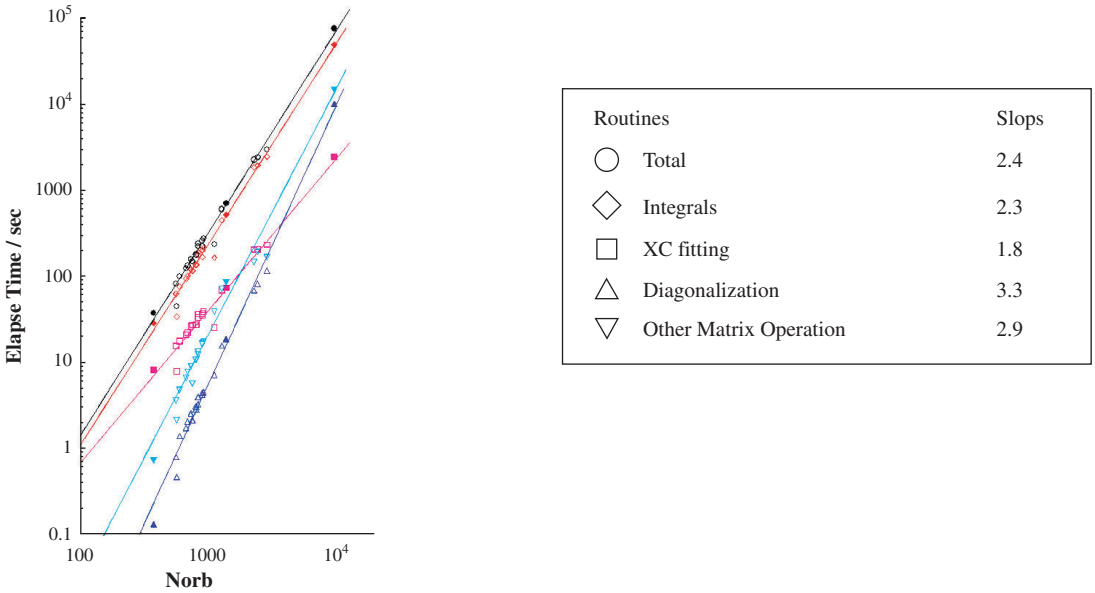

Fig. 2 Relation between Elapse Time of each Routine in SCF and the Number of Orbitals (Norb)

大規模計算によるタンパク質の動的構造および電子状態の研究

利用責任者
佐藤　文俊　　東京大学 生産技術研究所 客員助教授

著者
岡本　祐幸　　岡崎国立共同研究機構 分子科学研究所 助教授
斎藤　　稔　　弘前大学 理工学部 教授
佐藤　文俊　　東京大学 生産技術研究所 客員助教授

タンパク質は多数のアミノ酸残基がつながった巨大分子で、立体構造を形成し、時々刻々構造を変化させながら、電子を授受して固有の反応を進める。タンパク質の機能を理解するために、3つのサブグループによる研究を推進した。

A.「第一原理からのタンパク質の折れ畳みシミュレーション」
B.「タンパク質の高次構造変化のリアルなシミュレーション」
C.「密度汎関数法による超大型タンパク質の全電子計算」

今年度は、全てのサブグループで各々独自に開発したプログラムを地球シミュレータに移植し、チューニングに専念した。サブグループAのプログラムは、タンパク質の大域的安定構造を求めるREMDで、水中の小タンパク質の系で、ベクトル化率96.51%、並列化率92.88%、並列化効率80.98%を達成した。サブグループBのプログラムは、長距離クーロン力をカットオフせずにタンパク質が機能するときに起こす構造変化を観測するCOSMOS90で、98%のベクトル化率と8.5倍の加速を達成した。サブグループCのプログラムは、精密なタンパク質の全電子カノニカル軌道計算をおこなうProteinDFで、C++でコーディングされている。このコア部分のベクトル化に力を注ぎ、92%の効率を達成した。また、行列演算ルーチンは全てASL/ESに置き換えた。

次年度は、更なる高速化を達成し、具体的な系で計算を始める。

キーワード：タンパク質、折り畳み問題、拡張アンサンブル法、高次構造変化、分子動力学法、全電子計算、密度汎関数法